Discovery and Characterization of Novel Bioactive Peptides and a Natural ERRalpha Ligand

Discovery and Characterization of Novel Bioactive Peptides and a Natural ERRα Ligand

A dissertation presented

by

Adam Schwaid

to

The Department of Chemistry and Chemical Biology

in partial fulfillment of the requirements for the degree of Doctor of Philosophy

in the subject of

Chemistry

Harvard University

Cambridge, Massachusetts

August, 2013

Dissertation Advisor: Professor Alan Saghatelian Adam Schwaid

Discovery and Characterization of Novel Bioactive Peptides and a Natural ERRα Ligand

Abstract

Metabolites and peptides have a central role in biology that is often overlooked. Despite the importance of metabolites in key protein-metabolite interactions (PMIs), the extent and identity of these interactions is not known. Likewise, the extent to which short open reading frames (sORFs) in the genome are translated into peptides has also been an elusive question. This dissertation describes the development and application of methods to elucidate unknown molecules and interactions critical to understanding biology, and the subsequent characterization of the biological roles of these discoveries in cells and mice.

A liquid chromatography-mass spectrometry (LC-MS)-based metabolomics approach was used to discover that cholesterol is a ligand for Estrogen Related Receptor alpha (ERRα), an orphan nuclear receptor critical in numerous biological processes including metabolism, bone growth, and certain cancers. Despite intense study over the 25 years since this nuclear receptor was discovered there is no known endogenous ligand for this nuclear receptor. The discovery of cholesterol as a natural ERRα ligand allows for the understanding of how ERRα driven pathways are regulated and enables the modulation of receptor activity levels through control of cholesterol levels.

In addition, I also helped pioneer the development of a peptidomics strategy to discover novel sORF-encoded polypeptides (SEPs). Using our approach, we identified

86 novel SEPs. To further the discovery and characterization of these molecules, I collaborated on the development of a chemoproteomics approach to discover cysteine-containing SEPs (ccSEPs), leading to the identification of a further 17 SEPs. In total, 103 novel SEP, representing 103 novel human genes with unknown functions are now known.

To characterize SEPs, I developed a new workflow that relies on transcriptomics to characterize the functions of novel SEPs, and found that SEPs regulate gene expression. Based on changes in gene expression, SEPs can be assigned to several putative pathways. In one case, this analysis revealed that overexpression of the SEP results in a gene expression profile associated with addition of TNFα, which was confirmed by further biochemical characterization indicating this SEP promotes inflammation. More importantly, by establishing this approach, I have demonstrated a general strategy for elucidating the functions of SEPs.

Acknowledgements

I owe a great deal of thanks to Alan Saghatelian for his guidance and mentoring over the course of my Ph.D. If it were not for his willingness to take me under his wing I would never have had this opportunity to learn and grow as a scientist. His tutelage has allowed me to develop the skills to pose interesting scientific questions and answer them experimentally. Alan has fostered a uniquely collegial, collaborative, and friendly atmosphere in his laboratory that encourages the free flow of ideas and promotes teamwork amongst his students.

I would also like to thank my colleagues in the Saghatelian lab, from whom I learn every day. I would like to thank Sarah Slavoff for teaching me everything I know about experimental molecular biology, and frequent scientific conversations about varied topics—both relevant and irrelevant to our research. I would like to thank Andrew Mitchell who first mentored me when I joined the Saghatelian lab, and taught me how to ask the right questions when tackling mountains of information; Yui Vinayavekhin for her friendly welcome into the Saghatelian lab and her fastidious experimental advice; Anna Mari Lone for her tutelage in peptidomics, and good natured sense of humor; Amanda McFedries for her expertise in protein expression and all aspects of biochemistry—in addition to her keen insights into prospective business ventures; Tejia Zhang for her expertise in metabolomics and advice regarding experiments; Jiao Ma for her dedication, and quick mastery of new topics; Edwin Homan for his insights on mass spectrometry and science in general; Bogdan Budnik and John Neveau for advice on mass spectrometry and proteomics; Whitney Nolte for

frequent and unerring advice regarding graduate school and beyond, and Mathias Leidl for keeping lab lively and convivial.

Additionally, I would like to thank my colleagues in the Verdine Lab and Gregory Verdine for their counsel during the course of my Master's work.

The support of my family and friends was crucial to the completion of this dissertation. Their encouragement kept me going during the most challenging periods of my studies. In particular I would like to thank Steve Hershman for his prescient advice in the face of limited information. Most importantly, Julie French's unwavering faith and understanding in me formed the cornerstone of my strength without which my studies would have crumbled.

Table of Contents

List of Figures

List of Tables

Chapter 1: Methods for the Elucidation of Protein-Small Molecule Interactions

This chapter was adapted from:

McFedries A*, Schwaid A*, Saghatelian A. Methods for the Elucidation of Protein-Small Molecule Interactions. Chemistry and Biology, 2013, 20 (5), pp 667-673

*authors contributed equally.

1.1 Introduction

Understanding the interactions between small molecules and proteins can be approached from different perspectives and is important for the advancement of basic science and drug development. Chemists often use bioactive small molecules, such as natural products or synthetic compounds, as probes to identify therapeutically relevant protein targets. Biochemists and biologists often begin with a specific protein and seek to identify the endogenous metabolites that bind to it. These interests have led to the development of methodology that relies heavily on synthetic and analytical chemistry to identify protein-small molecule (PSMIs) and protein-metabolite interactions (PMIs). Here, we survey these strategies, highlighting key findings, to demonstrate the value of these approaches in answering important chemical and biological questions.

A number of different types of molecular interactions enable life. These include the interactions between proteins, proteins and nucleic acids, and proteins and small molecules. Elucidating these interactions and understanding how they control biology is a major scientific goal. A number of approaches have been developed in recent years to identify protein-protein interactions [1-6] and protein-nucleic acid interactions [7-9]. Ribosome profiling, for example, elucidates interactions between the ribosome and RNA in cells to reveal novel sites of protein translation. While many methods exist for the characterization of biopolymer interactions, far fewer approaches exist to elucidate interactions between proteins and small molecules. In recent years, however, more of these methods are beginning to emerge.

The importance of protein-small molecule interactions (PSMIs) in drug discovery and protein-metabolite interactions (PMIs) in biology has driven the development of new methods that rely heavily on the integration of both synthetic chemistry and analytical chemistry. Here, we divide these methods into two categories: small molecule-to-protein and protein-to-small molecule strategies. This division separates problems that aim to identify the protein targets of a bioactive small molecule from problems focused on identifying the small molecule-binding partner of a suspected metabolite-binding protein.

1.2 Small molecule-to-protein

1.2.1 Small-molecule affinity methods

One of the successes in using small molecules as affinity reagents is the identification of FKBP by the natural product FK506 [10]. This seminal work led to the discovery of new proteins and pathways that explained the mechanism of action of a potent class of immunosuppressant drugs [11]. In doing so, this work informed us about vital, but previously unknown, cellular pathways involved in the regulation of the cellular immune response. More generally, these studies highlighted the tremendous value of using bioactive small molecules to study biology. Other important examples, including the discovery of the histone deactylase family with trapoxin [12], reinforced the impact of chemistry in important biological discoveries, leading to the development of the field of chemical biology [13]. All of this is predicated on being able to use complex bioactive

small molecules as affinity reagents, which often requires complex chemical syntheses and emphasizes the importance of organic chemistry in this problem.

The increased use of small molecule screening approaches in biology has led to the identification of many bioactive molecules, leading to an increased demand for methods that can elucidate PSMIs. Affinity-based methods are still the most common approach used and leading methods have learned to integrate these approaches with modern proteomics to accelerate targeted discovery. Ong and colleagues, for example, have combined small molecule-affinity chromatography with SILAC, a quantitative mass spectrometry (MS)-based proteomics strategy, to identify PSMIs on a proteome-wide scale [14,15] (Figure 1.1). To demonstrate the generality of this approach seven different compounds, including kinase inhibitors and immunophilin ligands, were studied in this first example. Derivatives of each of these compounds were prepared and linked to solid support by a carbamate linkage, affording small molecule derivatized beads.

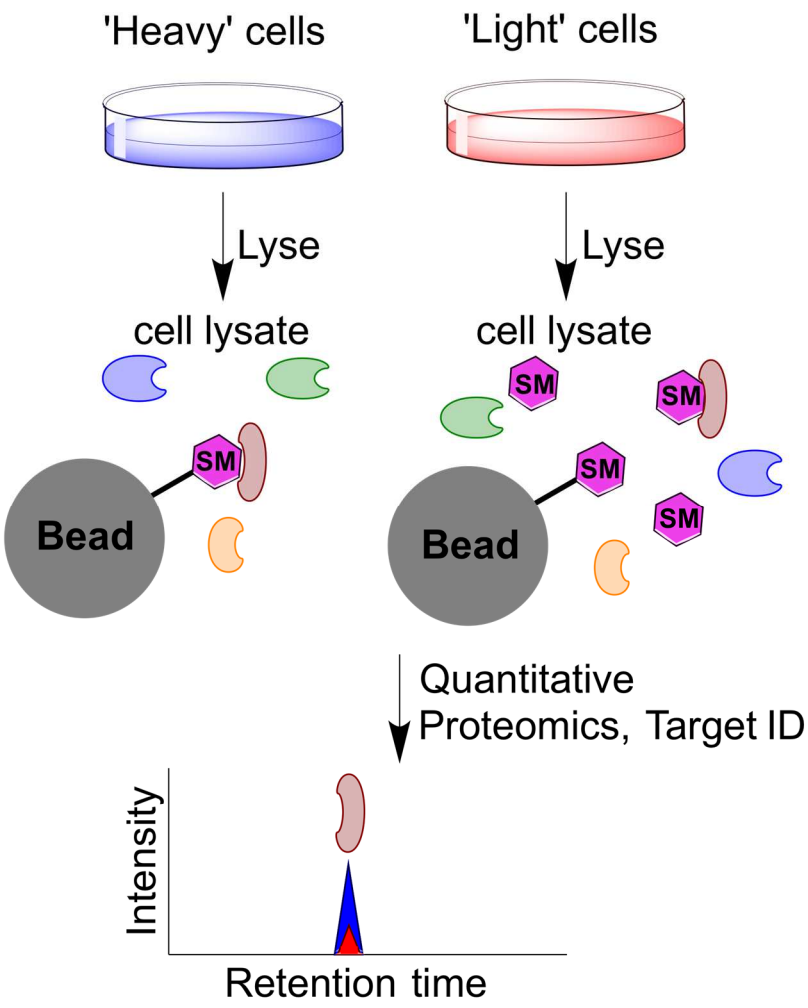


Figure 1.1. Affinity capture coupled to SILAC for small molecule target identification. Cells are grown in ‘heavy’ or ‘light’ media and subsequently lysed. These ‘heavy’ and ‘light’ cell lysates are separately incubated with beads that are modified with a small molecule of interest. Excess small molecule is added to the ‘light’ cell lysate and this soluble small molecule prevents any specific small molecule-protein interactions with the beads. After removal of the lysate bound proteins are eluted from the beads and the lysates are then combined for subsequent analysis using quantitative proteomics. The samples can be distinguished by mass spectrometry since proteins from each sample have different molecular weights, and therefore specific PMSIs can be identified by the higher concentrations of these ‘heavy’ proteins versus ‘light’ proteins.

In SILAC, cells are then grown in regular media (light) or specialized media (heavy) that replaces certain amino acids with stable isotope labeled derivatives (e.g. ^{13}C6-arginine and ^{13}C6,^{15}N2-lysine). The result is that the proteins in the ‘heavy’ cells have proteins that weigh more than the exact same proteins in ‘light’ cells, and these proteins can be distinguished and quantified by mass spectrometry. SILAC is used to identify PSMIs by passing ‘light’ lysate over beads coated with the bioactive small molecule. Any proteins with affinity for the small molecule are retained. As a control, a soluble variant of the small molecule is added to the ‘heavy’ lysate before it is incubated with the small molecule-modified beads. This soluble compound has the effect of binding of target proteins and preventing their binding to the small molecules on the surface of the bead. The post-bead lysate samples are then combined and analyzed by mass spectrometry. The ratio of light-to-heavy can identify those proteins that are specifically enriched by the small molecule on the bead, and therefore identify any target proteins of the small molecule.

The results from these initial experiments were excellent. All the compounds used identified known PSMIs, and several revealed some novel interactions. Moreover, by using compounds with different affinities for their targets, this work demonstrated that this method can successfully identify PSMIs with affinities from the low nanomolar (26 nM) to micromolar (44 μM) range. This strategy has been applied to the identification of the primary targets of piperlongumine [16], a compound that was shown to selectively kill cancer cells in vitro and in vivo by targeting the stress response to reactive oxygen species (ROS) [17]. Overall, these types of affinity approaches have come to dominate the methods that are used to identify PMSIs. Examples include the identification of the

nucleophosmin as a target of natural product avrainvillamide [18], the finding that cephalostatin A binds to specifically to members of the oxysterol binding-proteins, in the process revealing these proteins to be important in cancer cell proliferation[19].

Importantly, affinity methods are not limited to synthetic compounds or natural products, but can also be used with endogenous metabolites. Specifically, Nachtergaele and colleagues demonstrated a direct binding interaction between 20(S)-hydroxycholesterol and the oncoprotein smoothened (Smo), a key protein in the sonic hedgehog (Shh) pathway, using a derivative of 20(S)-hydroxycholesterol (nat-20(S)-yne) that was immobilized onto a solid support using Sharpless' click chemistry [20]. This example highlights the generality of affinity based approaches in identifying PSMIs. The only limitation appears to be the ability to access derivatives for immobilization by chemical synthesis that retain high affinity for their protein target. As scientists continue to gain interest in understanding the mechanism underlying bioactive small molecules, affinity-based methods will continue to be applied to many more target molecules.

1.2.1 Proteomic target identification

In cases where small molecules are difficult to modify, or the synthetic skill necessary to make such modifications are difficult to access, a new group of powerful proteomics methods to discover novel PSMIs can be used. These strategies rely on detecting differences in the stability between unbound and small-molecule bound proteins to identify the target(s) of a small molecule. Drug affinity responsive target stability (DARTS) is one such method [21]. DARTS relies on the fact that proteins are

more stable when bound to a metabolite, which makes them less susceptible to proteolysis. By comparing lysates with and without a small molecule in the presence of protease the small-molecule protein target(s) can be identified as the protein(s) that are more stable in the presence of the small molecule (Figure 1.2). Proof-of-concept experiments revealed that mammalian target of rapamycin (mTOR), for example, is less susceptible to proteolysis in the presence of E4, an mTOR kinase inhibitor. Moreover, DARTS is generally applicable and was used with other enzyme-inhibitor pairs, such as the COX2-celecoxib pair.

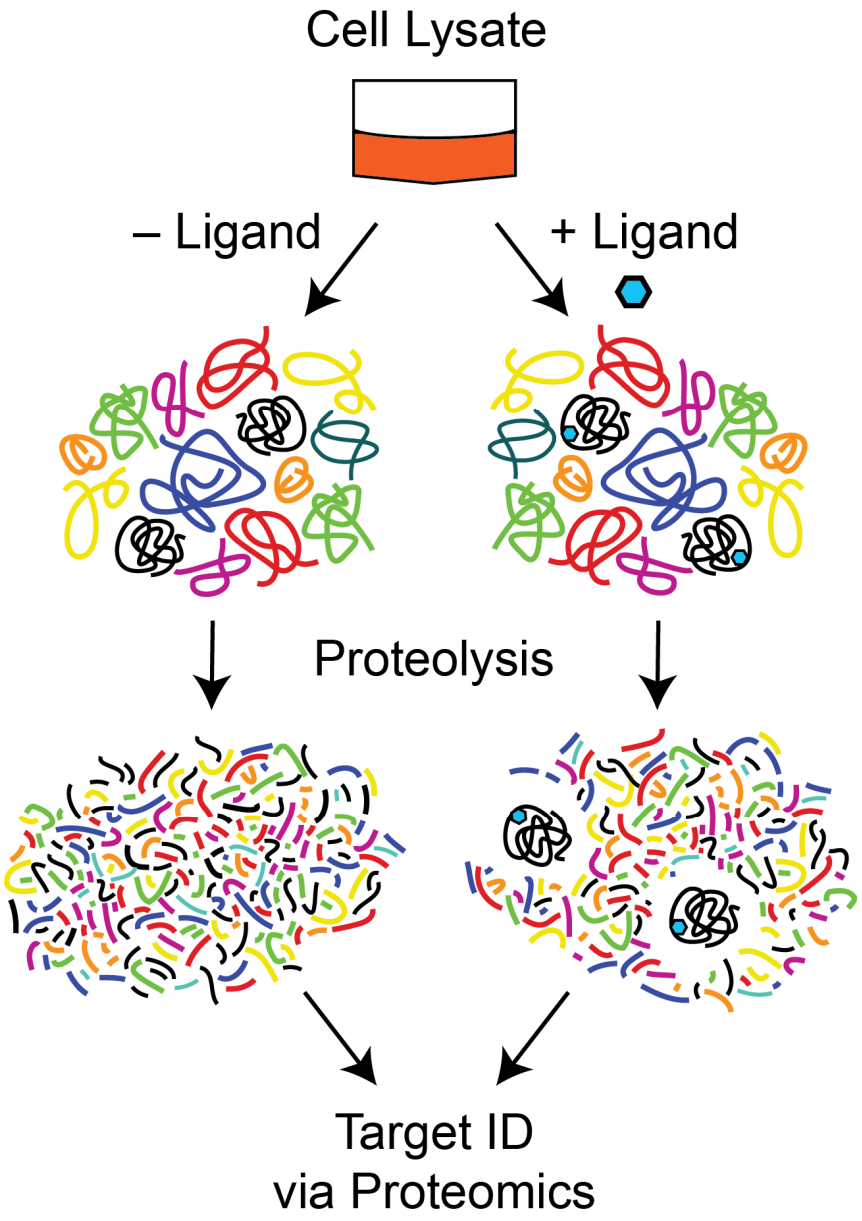


Figure 1.2. Drug Affinity Responsive Target Stability (DARTS). Aliquots of a protein lysate are mixed with either a small molecule (+ ligand) or solvent control (-ligand) to identify PSMIs. These samples are then subjected to limited proteolysis and compared by gel electrophoresis and quantitative mass spectrometry. Protein targets are identified as those proteins that display show increased protease resistance in the presence of the small molecule.

Next, DARTS was used to characterize PSMIs for resveratrol, an anti-aging compound thought to act primarily through interactions with the sirtuin protein Sirt1.

Using a yeast and human lysates the authors discovered that eukaroyotic initiation factor A (eIF4A) is a target of resveratrol, which demonstrates that DARTS is able to discover novel PSMIs. Importantly, the authors confirmed that at least some of the biology controlled by resveratrol is eIF4A dependent because worms lacking eIF4A no longer show any anti-aging in the presence of resveratrol. Together these experiments demonstrate the value and utility of DARTS for discovering PSMIs.

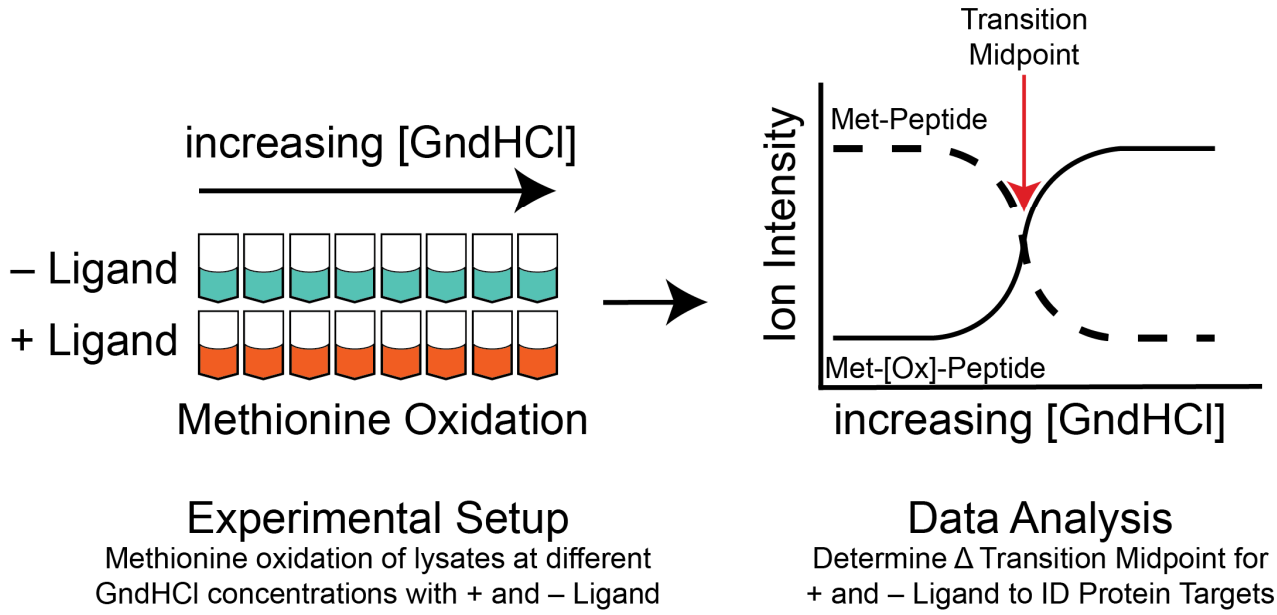


Figure 1.3. Stability of Proteins From Rates of Oxidation (SPROX). SPROx Identifies the targets of a small molecule from a complex protein mixture by measuring the ligand-induced changes in the rate of methionine amino acid side chain oxidation by hydrogen peroxide. Aliquots of the cell lysate are incubated with either a small molecule or a solvent control, then incubated with increasing concentrations of guanidine hydrocholride.
Ligand induced difference in target protein unfolding will impact the rate of selective methione oxidation by hydrogen peroxide. Quantitative proteomic is used to compare levels of nonoxidized versus oxidized methionine-containing peptides in each sample set (small molecule versus control) to determine the rate of target protein oxidation as a function of guanidine hydrocholride concentration. A shift of the transition midpoint for a protein between +ligand and –ligand samples indicates that the protein is a target for the small molecule in lysate.

Most recently, a proteomics method called Stability of Proteins from Rates of Oxidation (SPROX) was developed to identify PSMIs in complex cell lysates [22-25]. Rather than relying on non-specific proteolysis, which can make downstream mass spectrometry experiments difficult to interpret, SPROX relies on the irreversible oxidation of methionine residues by hydrogen peroxide to report on the thermodynamic stability of a protein's structure during chemical denaturation (Figure 1.3). Proof-of-

concept SPROX experiments performed using yeast lysates validated this approach by identifying known binders of the immunosuppressant drug cyclosporine A (CsA) [23]. SPROX has also been used with resveratrol, identifying the known target cytosolic aldehyde dehydrogenase, along with several novel interactions [25]. SPROX requires that target proteins contain methionine residues, and MudPIT analysis of SPROX experiments using yeast lysates demonstrated that 33% of the detectable proteins contain a methionine.

A SPROX-like method that uses s-methyl thioacetimidate (SMTA) labeling to detect amidination difference of proteins and protein-ligand complexes during chemical denaturation has also been explored [26]. This method requires target proteins to contain lysine residues or have a buried N-terminus in the native state. This method is particularly useful when a ligand of interest is not stable in hydrogen peroxide and therefore cannot be investigated by SPROX. SMTA labeling and SPROX complement each other, covering a wider range of the proteome [26]. Studying the thermodynamic stability of proteins under denaturing conditions in the presence and absence of ligand is an effective protein target identification strategy. However, the two described approaches require relatively larger concentrations of ligand, in the uM to mM range, but provide the flexibility to be used with a variety of downstream quantitative proteomic analysis.

1.2.1 Chemoproteomic target identification

Unfortunately, not all proteins are as active in lysates as they are within the context of a cell, and therefore some relevant PSMIs or PMIs may be missed when using lysates. Screening a spiroepoxide library for antiproliferative compounds, for example, revealed that the most relevant biological target of the active spiroepoxide is only targeted in a living cell, but was inactive once the cell was lysed [27]. Many molecular mechanisms can account for the difference between intact cells and lysates but the ultimate point is that it is difficult to exactly replicate cellular conditions in any type of biochemical experiment. Since it is impossible to predict what PSMIs are sensitive to cellular conditions new chemoproteomic approaches have been developed to enable target identification within live cells.

These methods rely on the use of small molecules that can covalently label their protein targets so that intracellular labeling events can be detected after cell lysis. Manabe and colleagues demonstrated the value of this approach with a modified natural product derivative in an effort to identify its protein target [28]. Potassium isolespedezate, a metabolite known to induce nyctinastic leaf opening in the motor cells of plants belonging to the Cassia genus, was derivitized with an iodoacetamide for covalent crosslinking to its target and an azide to enable enrichment and identification of the isolespedezate target by conjugation to a flag peptide using click chemistry. This approach led to the identification of 5-methyltetrahydropteroyltriglutamate-homocysteine S-methyltransferase (MetE) as the isolespedezate target protein. This method, while powerful, is mostly limited by the ability to synthesize natural product derivatives that can covalently label their target while maintaining the potency of the compound.

Most recently, a chemoproteomic strategy approach was developed for the identification of PMIs for the endogenous metabolite cholesterol. Cholesterol is a central metabolite with roles in membrane structure, metabolism, signaling and disease. While many important functions of this molecule are known, the full spectrum of proteins that interact with cholesterol is far from complete. Hulce and coworkers synthesized a series of cholesterol derivatives, and controls, containing a photocrosslinking diazirine group [29] (Figure 4). In practice, cells are irradiated by light after exposure to these sterol probes resulting in the covalent modification of any protein they bind. Addition of exogenous cholesterol blocks the overall labeling of these probes to validate that binding of these probes is occurring at cholesterol-specific binding sites. Subsequent to probe labeling, the probe is conjugated to biotin by CuACC chemistry allowing for affinity purification of labeled proteins.

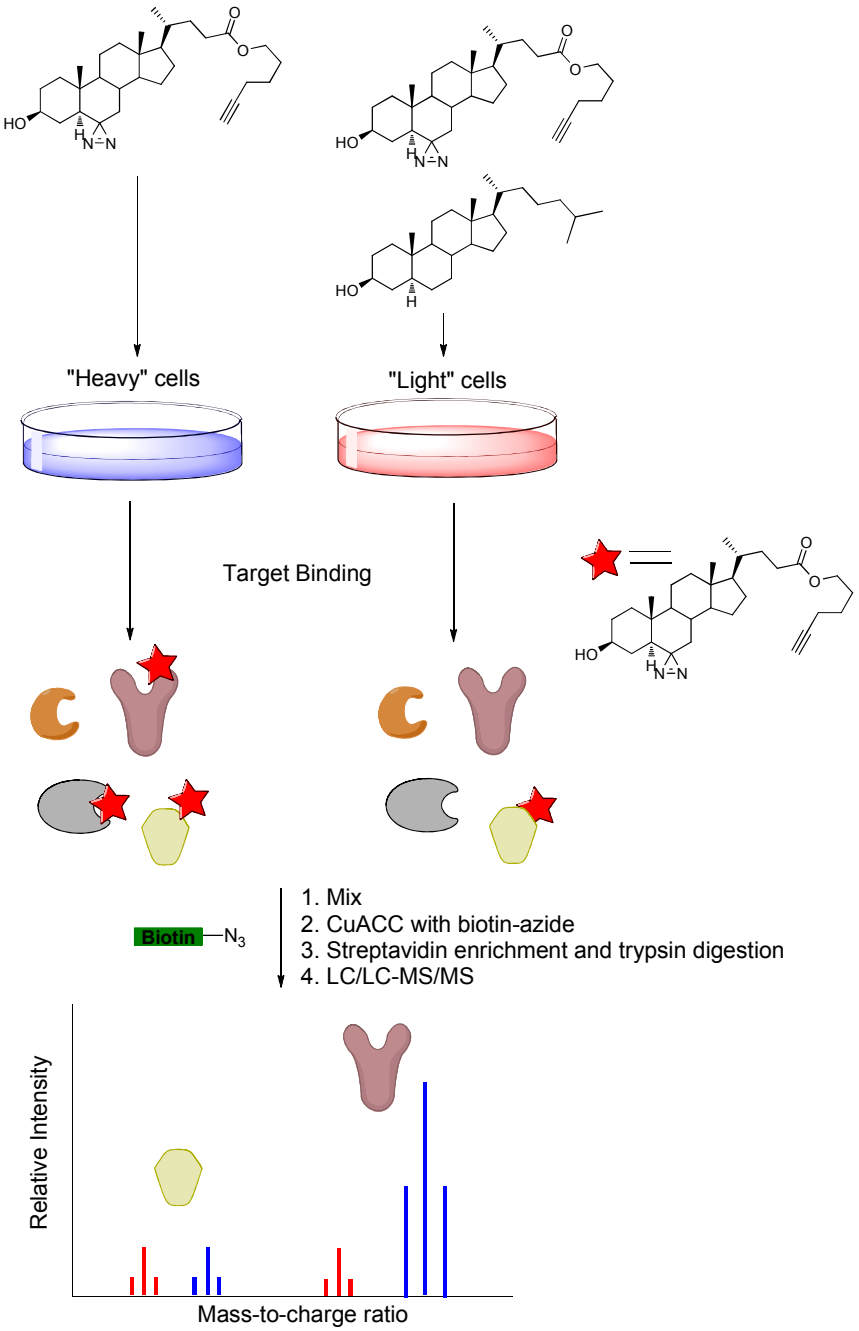


Figure 1.4: Identifying protein targets in cell culture. Cells grown in ‘heavy’ and ‘light’ media are treated with a cholesterol probe compound that has been modified to contain a diazirine moiety. In addition, excess cholesterol is also added to the ‘light’ sample and this will act to compete any specific cholesterol probe–protein interactions. The cholesterol probe is photocrosslinked to any bound protein targets in cells, and the cells are subsequently lysed. Any cholesterol probe–protein conjugates in this lysate are then modified with biotin using ‘click’ chemistry and labeled proteins are separated from cell lysate by affinity chromatography. The ‘heavy’ and ‘light’ fractions are then mixed and examined by quantitative proteomics. Proteins that specifically bind cholesterol will have a higher ratio of ‘heavy’ to ‘light’ in the mass spectrum.

Integrating these sterol probes with SILAC enables the identification of sterol probe-target proteins. In these experiments, the probe is added to both ‘heavy’ and ‘light’ cells, but the ‘light’ cells also contain a competitor (cholesterol) to block binding. Subsequent analysis of the ‘heavy’ and ‘light’ samples identifies cholesterol-binding proteins as those proteins enriched in the heavy sample versus the control sample. The identification of several known cholesterol-binding proteins such as Scap, caveolin and HMG-CoA reductase validated the methodology. Subsequent analysis of the entire data set using various bioinformatics tools revealed that almost every major class of protein has members that bind cholesterol, including GPCRs, ion channels and enzymes. More broadly, the analysis also revealed an enrichment of proteins involved in neurological disorders, cardiovascular and metabolic disease, demonstrating the potential therapeutic insights that may eventually be provided by this data.

Together these examples demonstrate the utility of chemoproteomic approaches to identify PSMIs and PMIs, and highlight the power of these approaches to rapidly increase our understanding of the role of specific small molecules in biology.

1.3 Protein-to-small molecule

1.3.1 Biophysical identification of small molecule binders

In many cases, the problem of identifying a PMI begins with interest in a particular protein. This protein may be a potential drug target or it may be suspected to require small molecule binding to regulate its activity. There are numerous cell-based

assays for GPCRs, nuclear receptors, enzymes and many more proteins that are routinely used to identify new small molecule ligands. In general, these methods are highly effective. Such assays have already been reviewed extensively in the literature. Instead, this chapter focuses on cell-free approaches that rely heavily on biochemical, biophysical and profiling methods to reveal PMIs for endogenous metabolites.

Biophysical screening methods provide an effective means for PMI discovery from endogenous metabolites. Differential scanning techniques were originally developed to optimize recombinant protein stability (i.e. melting temperature (Tm)) for purification and crystallography [30]. Differential static light scattering (DSLS) and/or differential scanning fluorometry (DSF) are the two most commonly used methods. DSLS measured denaturation by tracking temperature induced increases in the intensity of scattered light, while DSF measured increases in the fluorescence from the environmentally sensitive dye SYPRO Orange. Shifts in melting temperatures of > 2 °C were found to confidently represent binding events and conditions that enhanced protein stability, and thermal shifts > 4 °C increased the likelihood of positive results in crystallographic screens.

These methods have recently been extended to identify PMIs by measuring the effect of small molecules on the melting temperature (Tm) of proteins. The binding between a small molecule and protein stabilizes the protein structure (i.e. raises the Tm) (Figure 1.5). Recently, DeSantis and coworkers used DSF to identify natural estrogen receptor alpha (ERα) from a library of molecules [31]. DSF successfully identified known natural ERα agonists, β-estradiol and estrone, demonstrating the utility of this assay in characterizing natural PMIs. The authors suggest that these assays will be useful in the

identification of unknown nuclear receptor ligands in the future.

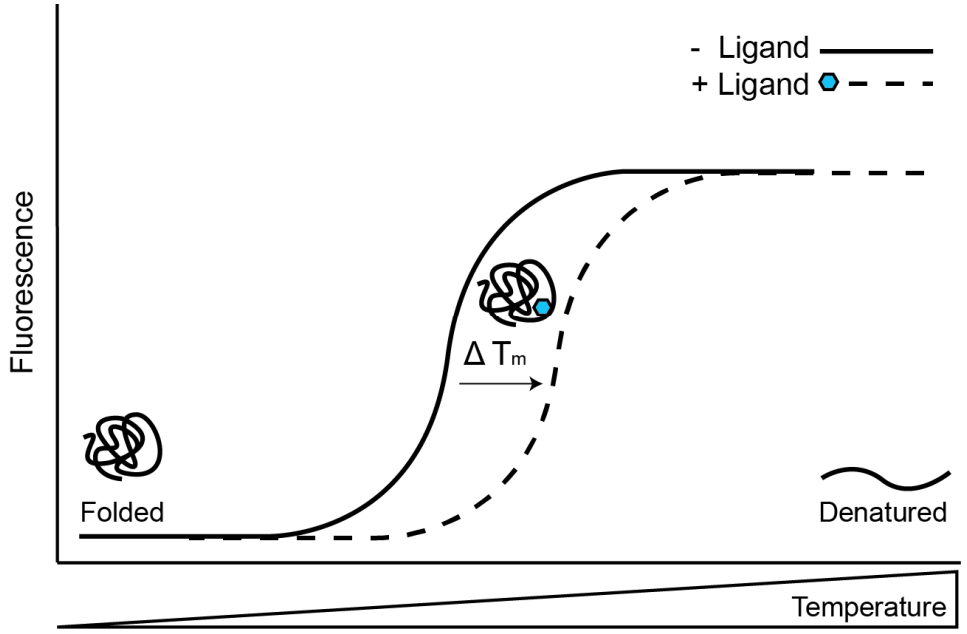


Figure 1.5: Thermostability Shift Assay. PSMI and PMIs can be identified in a high-throughput fashion by monitoring shifts in the melting temperature (Tm) of a protein-ligand complex versus a protein-solvent control. The environmentally sensitive dye SYPRO Orange emits when bound to hydrophobic amino acid residues and is used to monitor protein unfolding via differential scanning fluorimetry (DSF).

1.3.2 Affinity-based identification of small molecule binders

Alternatively, affinity based experiments using immobilized proteins are another option for the characterization of PMIs. In general, these assays provide the advantage that they can be performed with unmodified metabolite resulting in a reduced likelihood of false negatives. This approach relies on the fact that proteins can be immobilized without altering the secondary structure of the protein or interfering with ligand binding. The first examples of affinity methods relied on using radioisotopes to identify PMIs by

measuring radioactivity of a protein after incubation and washing with a radiolabeled ligand [32,33].

Most recently, this approach has been optimized in the form of a DRaCALA assay (Differential radial capillary action of ligand assay), which allows for the rapid high throughput identification of PMIs using radiolabelled metabolites [34]. DRaCALA utilizes the affinity of proteins for nitrocellulose membranes to sequester radiolabelled metabolites bound to protein. A solution containing the protein of interest and radiolabelled ligand is spotted on nitrocellulose. Unbound ligand will diffuse with the solvent throughout the membrane, whereas protein and ligand bound to protein will be immobilized at the point it was spotted. One advantage of DRaCALA is that it can be performed using whole cell lysates to identify PMIs, which avoids time-consuming protein purification.

The one disadvantage of this method is that it requires foreknowledge as to what candidate metabolites should be tested, and in that each metabolite must be tested individually. Nevertheless, for certain cases DRaCALA provides a high-throughput means to identify ligand-binding proteins. The identification of prokaryotes that have the proteins capable of binding bis-(3'-5')-cyclic dimeric guanosine monophosphate (cdiGMP), a metabolite important in biofilm formation, was accomplished by simply using lysates from 191 strains of P. aeruginosa and 82 other bacterial strains. The 'hits' in this assay corresponded to those bacteria that have diguanylate cyclase (DGC), as expected. This approach precludes the identification of unexpected PMIs or PMIs with novel metabolites. Still, DRaCALA remains a powerful approach for uncovering protein metabolite interactions.

The use of global mass spectrometry approaches allows for the scrutiny of a larger pool of metabolites, including novel metabolites, and can result in the unbiased identification of protein metabolite binding. Specifically, the use of global metabolite profiling enabled the development of a novel, unbiased, strategy for the identification of endogenous PMIs [35] (Figure 1.6). In this approach, recombinant proteins fused to an affinity tag—either GST or hexahistidine—are immobilized on a solid support. Incubation of these proteins with a metabolite mixture, typically an extract containing the entire lipidome from a cell or tissue of interest results in the formation of a protein-metabolite complex on the bead. Following this incubation, the protein is washed, subsequently eluted from the solid support, and the eluant is then analyzed using a liquid chromatography–mass spectrometry (LC–MS) metabolite profiling platform. Quantitative comparison of the metabolite profiles between samples with and without protein reveals any metabolites that are enriched by the protein.

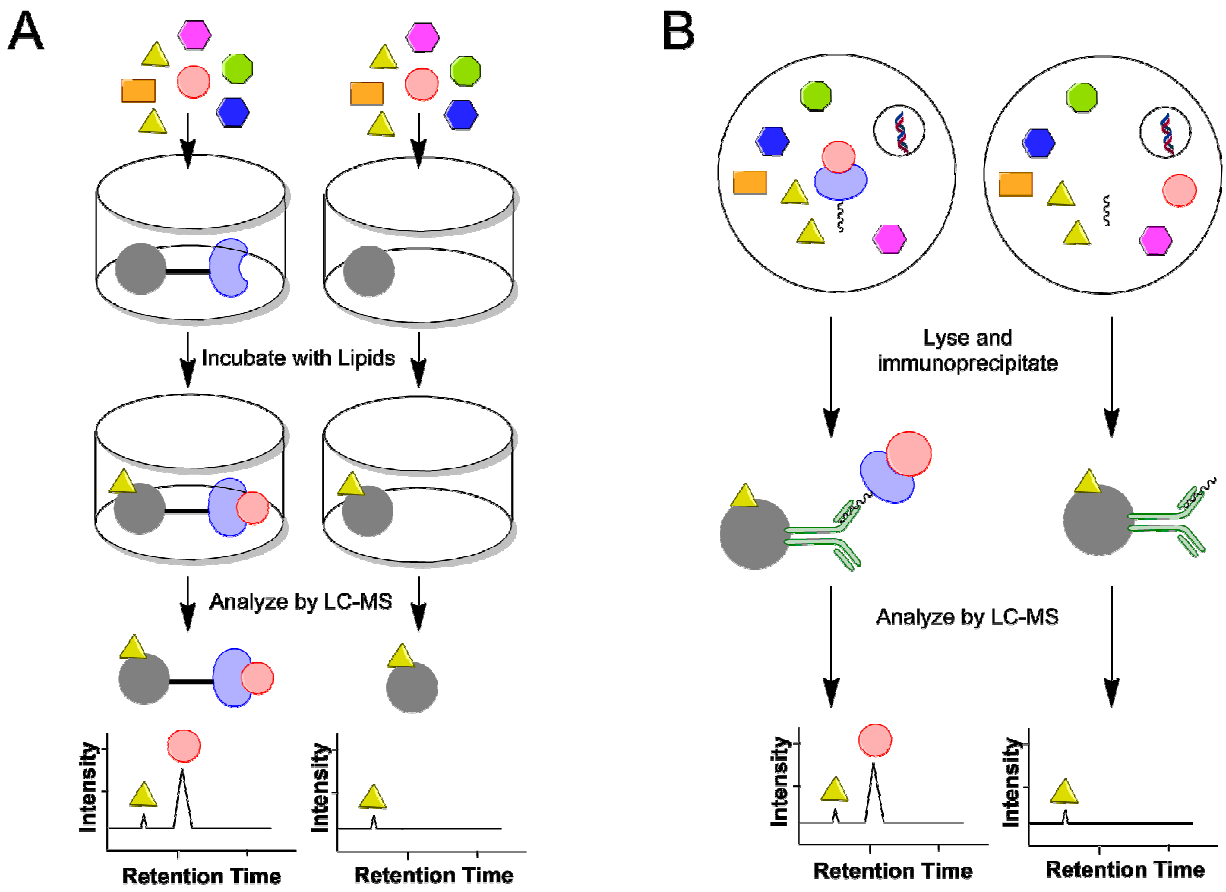


Figure 1.6. Affinity methods for elucidating PMIs. A) Protein (blue) is immobilized on a solid support (grey) and incubated with lipids from tissue lysate. The immobilized protein is then washed and the protein is eluted from the beads. The eluate is then analyzed by LC-MS and compared to eluate from a control sample (solid support with no protein) by in order to identify specific PMIs. B) A yeast strain bearing either the protein of interest fused to an IgG epitope tag, or the IgG epitope tag only, were lysed and immunoprecipitated using antibody labeled beads. Lipids were then extracted from the beads and examined by LC-MS. Comparison between the protein and control sample can be used to identify any specific PMIs.

This approach was developed using three different lipid binding proteins with known ligands: cytosolic retinoic acid binding protein 2 (CRABP2), fatty acid binding protein 2 (FABP2) and StarD3. The strategy successfully identified specific PMIs for all three of these proteins, and in doing so created a reliable method to pulldown protein

metabolite interactions [35]. It does not require any explicit knowledge about the identity of the metabolite and can be conducted rapidly against a large pool of metabolites. Moreover, it was not susceptible to the enrichment of nonspecific metabolites. Although the examples presented here centered around lipids, there is no reason this approach could not be used for polar metabolites as well. This strategy has been successfully applied to other PMIs, including the discovery that arachidonic acid and docosahexanoic acid, polyunsaturated fatty acids, bind to the orphan nuclear receptor Nur77 [36].

Li and colleagues developed an exciting strategy to identify PMIs within yeast [37] (Figure 1.6). Specifically, they focused on interaction with yeast protein kinsases as well as proteins that comprise ergosterol pathway. The proteins in this group were epitope tagged to enable their immunoprecipitation from cellular lysates. These immunoprecipitated samples were analyzed by metabolomics to identify any endogenous metabolites bound to the proteins. As a control, each sample was also checked by SDS-PAGE to ensure pulldown of the target protein was successful.

Comparison of the metabolite profiles from these various samples revealed a number of new interactions. Many of the enzymes within the ergosterol pathway bound to sterol intermediates, which suggest that these molecules exert regulation on the pathway. Interestingly, it was also discovered that some of the proteins within the ergosterol pathway bind to pentaporphoryin, which was a complete surprise but helps explain a previous observation linking pentaporphoryin and ergosterol regulation. Furthermore, their analysis led to the discovery that a number of yeast protein kinases are regulated by ergosterol highlighting the generality of this method towards numerous

protein classes. In aggregate, this data highlights the potential for unbiased PMI identification to greatly increase our current understanding of endogenous small molecule biology.

1.4 Conclusions

A successful library of approaches to determine PSMIs and PMIs has been developed, and is ready to be applied to identify unknown PMIs of interest. These approaches enable the discovery of novel interactions and are also designed to maximize the likelihood that the interactions are occurring in cells and tissues. The continued application of these methods will enrich our understanding of small molecule biology and also stimulate the development of improved methods for discovering these interactions. As demonstrated by the above examples this research area sits squarely at the interface of chemistry and biology and will greatly benefit from collaboration between future generations of chemists, biochemists, and biologists.

1.5 References

(1) Gingras, A. C.; Gstaiger, M.; Raught, B.; Aebersold, R. *Nature reviews. Molecular cell biology* **2007**, *8*, 645.

(2) Uetz, P.; Giot, L.; Cagney, G.; Mansfield, T. A.; Judson, R. S.; Knight, J. R.; Lockshon, D.; Narayan, V.; Srinivasan, M.; Pochart, P.; Qureshi-Emili, A.; Li, Y.; Godwin, B.; Conover, D.; Kalbfleisch, T.; Vijayadamodar, G.; Yang, M.; Johnston, M.; Fields, S.; Rothberg, J. M. *Nature* **2000**, *403*, 623.

(3) Frei, A. P.; Jeon, O. Y.; Kilcher, S.; Moest, H.; Henning, L. M.; Jost, C.; Pluckthun, A.; Mercer, J.; Aebersold, R.; Carreira, E. M.; Wollscheid, B. *Nature biotechnology* **2012**, *30*, 997.

(4) Kaushansky, A.; Allen, J. E.; Gordus, A.; Stiffler, M. A.; Karp, E. S.; Chang, B. H.; MacBeath, G. *Nature protocols* **2010**, *5*, 773.

(5) Hubner, N. C.; Bird, A. W.; Cox, J.; Splettstoesser, B.; Bandilla, P.; Poser, I.; Hyman, A.; Mann, M. *The Journal of cell biology* **2010**, *189*, 739.

(6) Glatter, T.; Wepf, A.; Aebersold, R.; Gstaiger, M. *Molecular systems biology* **2009**, *5*, 237.

(7) Ingolia, N. T.; Brar, G. A.; Rouskin, S.; McGeachy, A. M.; Weissman, J. S. *Nature protocols* **2012**, *7*, 1534.

(8) Johnson, D. S.; Mortazavi, A.; Myers, R. M.; Wold, B. *Science* **2007**, *316*, 1497.

(9) Ingolia, N. T.; Ghaemmaghami, S.; Newman, J. R. S.; Weissman, J. S. *Science* **2009**, *324*, 218.

(10) Harding, M. W.; Galat, A.; Uehling, D. E.; Schreiber, S. L. *Nature* **1989**, *341*, 758.

(11) Schreiber, S. L. C., G.R. *Immunology Today* **1992**, *13*, 136.

(12) Taunton, J.; Collins, J. L.; Schreiber, S. L. *Journal of the American Chemical Society* **1996**, *118*, 10412.

(13) Kijima, M.; Yoshida, M.; Sugita, K.; Horinouchi, S.; Beppu, T. *Journal of Biological Chemistry* **1993**, *268*, 22429.

(14) Ong, S.-E.; Blagoev, B.; Kratchmarova, I.; Kristensen, D. B.; Steen, H.; Pandey, A.; Mann, M. *Molecular & Cellular Proteomics* **2002**, *1*, 376.

(15) Ong, S.-E.; Schenone, M.; Margolin, A. A.; Li, X.; Do, K.; Doud, M. K.; Mani, D. R.; Kuai, L.; Wang, X.; Wood, J. L.; Tolliday, N. J.; Koehler, A. N.; Marcaurelle, L. A.; Golub, T. R.; Gould, R. J.; Schreiber, S. L.; Carr, S. A. **2009**, *106*, 4617.

(16) Raj, L.; Ide, T.; Gurkar, A. U.; Foley, M.; Schenone, M.; Li, X.; Tolliday, N. J.; Golub, T. R.; Carr, S. A.; Shamji, A. F.; Stern, A. M.; Mandinova, A.; Schreiber, S. L.; Lee, S. W. *Nature* **2011**, *475*, 231.

(17) Adams, D. J.; Dai, M.; Pellegrino, G.; Wagner, B. K.; Stern, A. M.; Shamji, A. F.; Schreiber, S. L. *Proc. Natl. Acad Sci.* **2012**, *109*, 15115.

(18) Wulff, J. E.; Siegrist, R.; Myers, A. G. *Journal of the American Chemical Society* **2007**, *129*, 14444.

(19) Burgett, A. W. G.; Poulsen, T. B.; Wangkanont, K.; Anderson, D. R.; Kikuchi, C.; Shimada, K.; Okubo, S.; Fortner, K.; Mimaki, Y.; Kuroda, M.; Murphy, J. P.; Schwalb, D. J.; Petrella, E. C.; Cornella-Taracido, I.; Schirle, M.; Tallarico, J. A.; Shair, M. D. *Nature Chemical Biology* **2011**, *7*, 639.

(20) Nachtergaele, S.; Mydock, L. K.; Krishnan, K.; Rammohan, J.; Schlesinger, P. H.; Covey, D. F.; Rohatgi, R. **2012**, *8*, 211.

(21) Lomenick, B.; Hao, R.; Jonai, N.; Chin, R. M.; Aghajan, M.; Warburton, S.; Wang, J.; Wu, R. P.; Gomez, F.; Loo, J. A.; Wohlschlegel, J. A.; Vondriska, T. M.; Pelletier, J.; Herschman, H. R.; Clardy, J.; Clarke, C. F.; Huang, J. *Proceedings of the National Academy of Sciences* **2009**, *106*, 21984.

(22) West, G. M.; Tang, L.; Fitzgerald, M. C. *Analytical Chemistry* **2008**, *80*, 4175.

(23) West, G. M.; Tucker, C. L.; Xu, T.; Park, S. K.; Han, X.; Yates, J. R.; Fitzgerald, M. C. *Proceedings of the National Academy of Sciences* **2010**, *107*, 9078.

(24) Strickland, E. C.; Geer, M. A.; Tran, D. T.; Adhikari, J.; West, G. M.; DeArmond, P. D.; Xu, Y.; Fitzgerald, M. C. *Nat. Protocols* **2013**, *8*, 148.

(25) DeArmond, P. D.; Xu, Y.; Strickland, E. C.; Daniels, K. G.; Fitzgerald, M. C. *Journal of Proteome Research* **2011**, *10*, 4948.

(26) Xu, Y.; Falk, I. N.; Hallen, M. A.; Fitzgerald, M. C. *Analytical Chemistry* **2011**, *83*, 3555.

(27) Evans, M. J. S., Alan; Jorenson, Erik J.; Cravatt, Benjamin F. *Nature biotechnology* **2005**, *23*, 1303.

(28) Manabe, Y. M., Makoto; Ito, Satoko; Kato, Nobuki; Ueda, Minoru *Chemical Communications* **2010**, *46*, 469.

(29) Hulce, J. J.; Cognetta, A. B.; Niphakis, M. J.; Tully, S. E.; Cravatt, B. F. *Nature methods* **2013**, *10*, 259.

(30) Vedadi, M.; Niesen, F. H.; Allali-Hassani, A.; Fedorov, O. Y.; Finerty, P. J.; Wasney, G. A.; Yeung, R.; Arrowsmith, C.; Ball, L. J.; Berglund, H.; Hui, R.; Marsden, B. D.; Nordlund, P.; Sundstrom, M.; Weigelt, J.; Edwards, A. M. *Proceedings of the National Academy of Sciences* **2006**, *103*, 15835.

(31) DeSantis, K. *Nuclear Receptor Signaling* **2012**, *10*.

(32) Sundberg, S. A. *Current Opinions in Biotechnology* **2000**, *11*, 47.

(33) Bosworth, N. T., P.; *Nature* **1989**, *341*, 167.

(34) Roelofs, K. G. W., Jingxin; Sintim, Herman O.; Lee, Vincent T. *Proc. Natl. Acad Sci.* **2011**, *108*, 15528.

(35) Tagore, R. T., Horatio R.; Homan, Edwin A.; Munawar, Ali; Saghatelian, Alan *Journal of the American Chemical Society* **2008**, *130*, 14111.

(36) Vinayavekhin, N. S., Alan *Journal of the American Chemical Society* **2011**, *133*, 17168.

(37) Li, X.; Gianoulis, T. A.; Yip, K. Y.; Gerstein, M.; Snyder, M. **2010**, *143*, 639.

Chapter 2: Cholesterol is a Natural ERRα Ligand

This chapter was adapted from:

Schwaid, AG*, Wei W*, Wang X, Saghatelian A, Wan Y. Cholesterol is a Natural ERRα Ligand. Submitted

*authors contributed equally

2.1 Introduction

Estrogen-related receptors are a family of orphan nuclear receptors that consist of ERRα, ERRβ and ERRγ[1,2]. Through their regulation of transcription, these proteins control a variety of physiological and pathological pathways. ERRα is the most studied member of this nuclear receptor subgroup, yet has eluded deorphanization for over a quarter of a century. ERRα is a critical regulator of bone remodeling by controlling key bone cell differentiation processes such as osteoclastogenesis; ERRα deletion attenuates osteoclast differentiation and bone resorption leading to increased bone mass[3]. ERRα also modulates energy metabolism by controlling adipogenesis, lipogenesis, insulin sensitivity, mitochondria biogenesis and fatty acid oxidation; ERRα deletion or inhibition confers resistance to obesity and insulin resistance[4-6]. Furthermore, clinical and basic research has revealed ERRα as an important regulator of multiple cancers[4,7]. Despite the centrality of ERRα in human biology, an endogenous ligand for ERRα has been elusive.

ERRα was originally identified based on its homology to the estrogen receptor α (ERα)[2]. Analysis of the individual domains reveals that the DNA-binding domains (DBDs) of ERRα and ERα are 70% homologous, but their ligand-binding domains (LBDs) are only 36% similar. Indeed, while ERα transcriptional activity is regulated by 17β-estradiol, estrone or estriol binding to its LBD[8], these steroid hormones have no impact on ERRα function[9]. Nuclear receptor transcriptional activity is usually regulated by metabolite ligands; however ERRα is constitutively active in the absence of an exogenous small molecule. Moreover, the ligand-binding pocket of ERRα bound to a co-

activator peptide is almost completely occluded by hydrophobic residues. Therefore, the current hypothesis is that ERRα is a ligand-independent nuclear receptor[10].

The importance of ERRα biology in human health has led to tremendous interest in this protein as a novel therapeutic target. A number of synthetic small-molecule ERRα antagonists have been developed. These compounds bind to the ERRα ligand-binding pocket and induce a conformational shift in the LBD that interferes with co-activator binding and inactivates transcription. ERRα antagonists have been found to induce cancer cell death[11], inhibit tumor growth[12], and improve insulin sensitivity and glucose tolerance[5]. More generally, the discovery of these antagonists indicates that ERRα has a functional small-molecule binding pocket, renewing the idea that ERRα has an endogenous ligand.

2.2 Discovery of an endogenous ERRα Binder

The identification of endogenous nuclear receptor ligands is typically accomplished using functional screens to identify metabolites that regulate nuclear receptor transcriptional activity. Screening several phospholipids in a transcriptional reporter assay, for example, has led to the identification of an endogenous LRH-1 ligand[13]. While successful in some cases, cell-based assays with natural metabolites have certain drawbacks that could potentially lead to false negatives. Recent work, for example, has demonstrated the importance of cellular proteins in the transport[14] and metabolism[15] of endogenous ligands. To obviate the need to consider these variables, we decided on an unbiased metabolite profiling strategy to identify an endogenous ligand for ERRα.

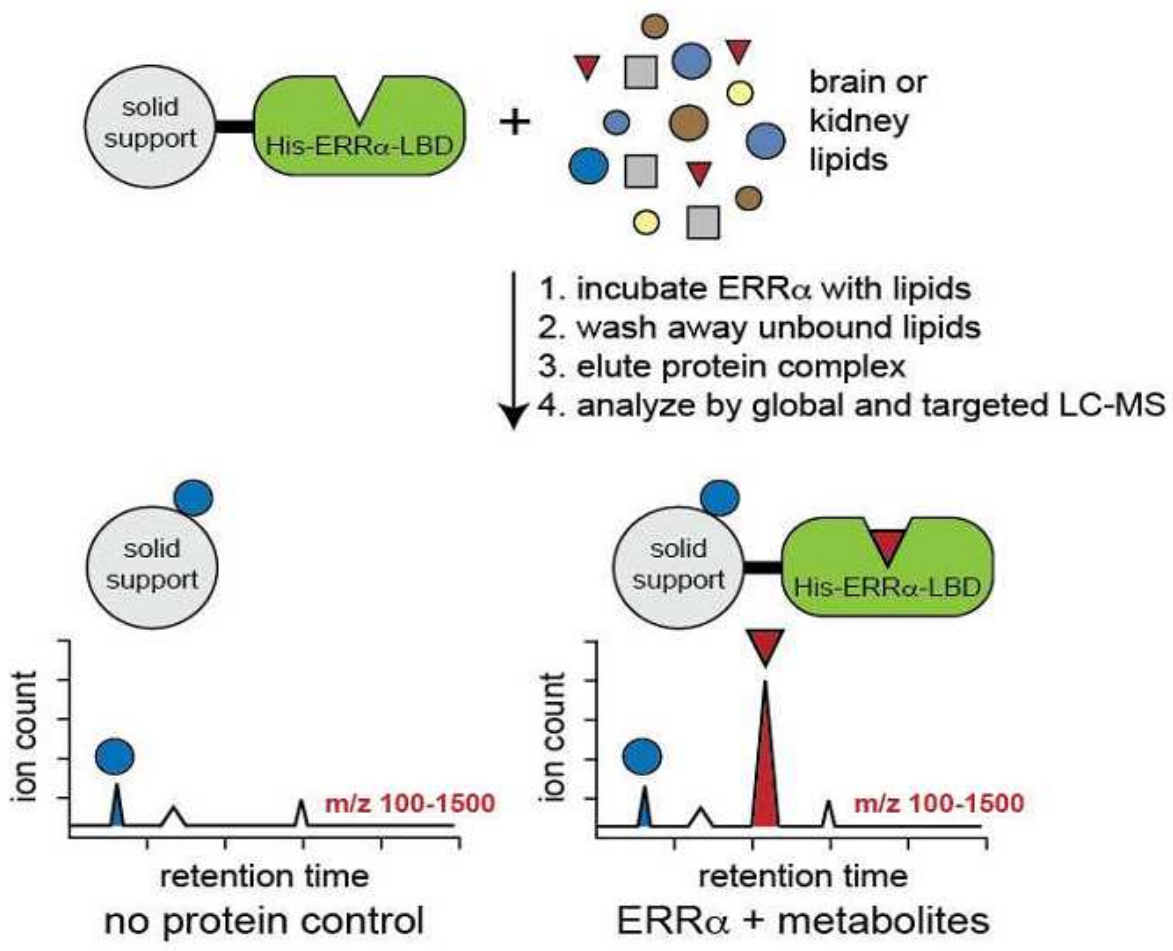


Figure 2.1: Metabolomics approach to identify ERRα binders. His-ERRα-LBD was immobilized on a solid support, and incubated with lipids from brain or kidney. After incubation, unbound lipids were washed away and protein was eluted. Eluant was then analyzed by LC-MS TOF and compared to a no protein control. Differences in detected lipids can be attributed to the protein binding.

In this approach, comparative metabolite profiling is used to identify metabolites that are specifically enriched from a cellular extract by resin bound ERRα-LBD (Fig. 2.1). We previously used this method to enrich endogenous PPARγ ligands from cells[16], indicating the compatibility of this approach with nuclear receptors. Recombinant hexahistidine-tagged ERRα-LBD (HIS-ERRα-LBD) was expressed, purified and immobilized on a nickel bound resin. Resin loaded with ERRα-LBD was used to enrich

lipids from brain or kidney extracts, which were selected because of the robust expression of ERRα in these tissues. Analysis of the metabolite profiling data from this experiment compared to a control (unloaded resin) identified cholesterol as the only metabolite that was significantly enriched ($p<0.05$ and >2-fold) by the HIS-ERRα-LBD resin, revealing a specific binding interaction between ERRα-LBD and cholesterol (Figure 2.2).

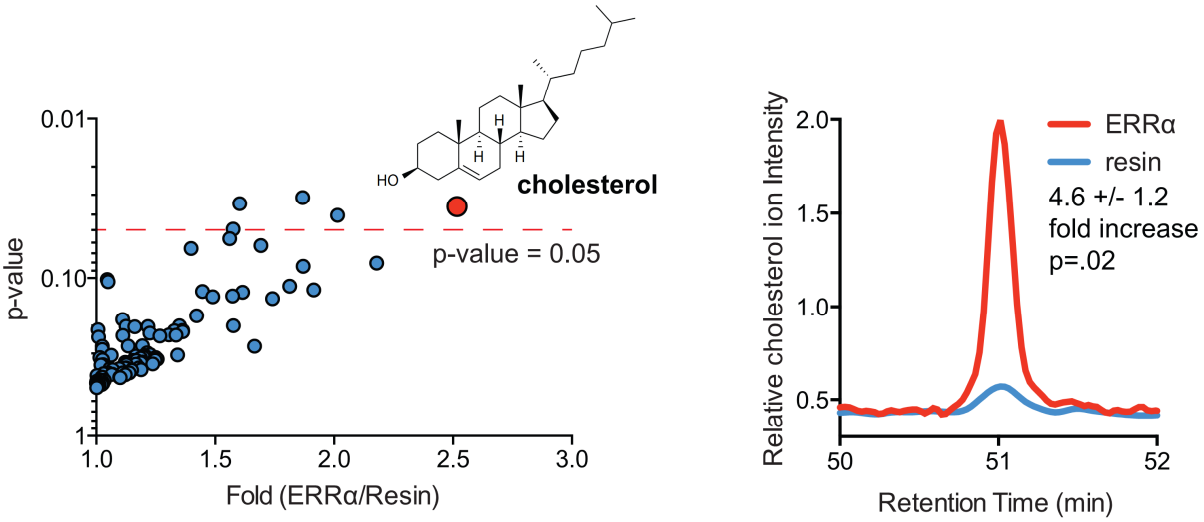


Figure 2.2: ERRα binds to cholesterol. (Left) Using a global lipid pulldown from brain lysate, cholesterol was identified as the only statistically enriched binder. (Right) In order to determine whether cholesterol binding was specific or whether other sterols were enriched by ERRα, immobilized ERRα was incubated with sterols from brain or kidney, washed, and protein was eluted. Lipids were analyzed by a targeted MRM method in order to sensitively identify ERRα binders. Again, cholesterol was the only sterol found to bind ERRα indicating cholesterol ERRα binding is specific. This result could be repeated from brain or kidney lysate. Statistics was performed with a Student's t-test, *, $p \leq 0.05$

Several further experiments were performed to validate this interaction. First, the addition of the synthetic ERRα antagonist diethylstilbestrol (DES) to tissue extracts inhibited cholesterol enrichment by the HIS-ERRα-LBD resin (Figure 2.3). Metabolite profiling of this experiment revealed that HIS-ERRα-LBD enriched DES, implying that DES binding prevented cholesterol binding. Next, the addition of the synthetic ERRα

antagonist/inverse agonist XCT790, DES, or cholesterol to the ERRα LBD all resulted in a conformational change in the protein whereas the negative control compound estradiol did not (Figure 2.4). Lastly, a tryptophan fluorescence assay indicated that cholesterol bound to the HIS-ERRα-LBD with a kD of ~700 nM. Estradiol had no effect on HIS-ERRα-LBD tryptophan fluorescence, indicating specificity in ERRα-cholesterol interaction (Figure 2.5). These data validate the existence of a specific and high affinity ERRα-cholesterol binding interaction.

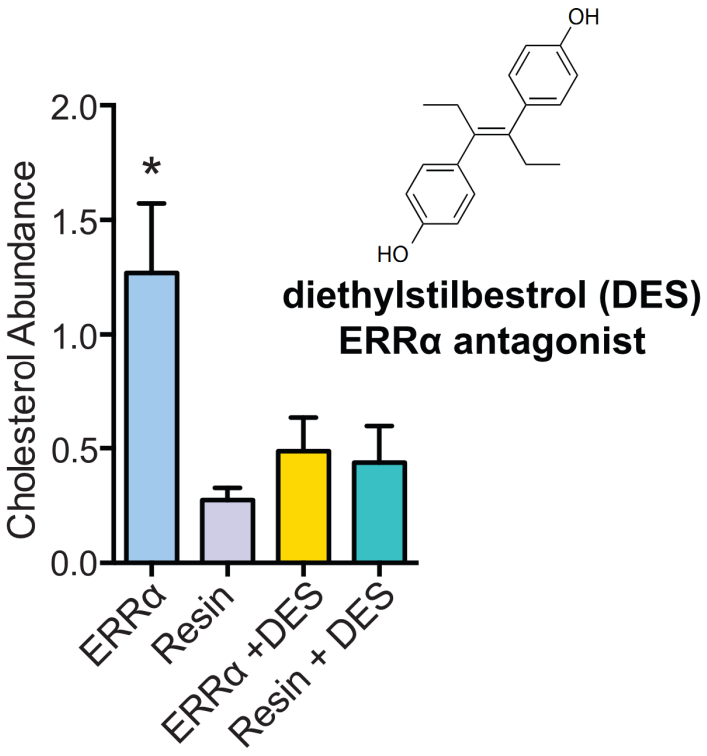


Figure 2.3: Diethylstilbestrol blocks ERRα cholesterol binding. Brain lipids were spiked with DES (79uM), and this lipid mixture was incubated with immobilized ERRα. After washing eluant was analyzed by mass spectrometry. DES binding to ERRα was observed by LC-MS TOF, and cholesterol binding was blocked in the presence of DES. Statistics was performed with a Student's t-test, *, $p \leq 0.05$

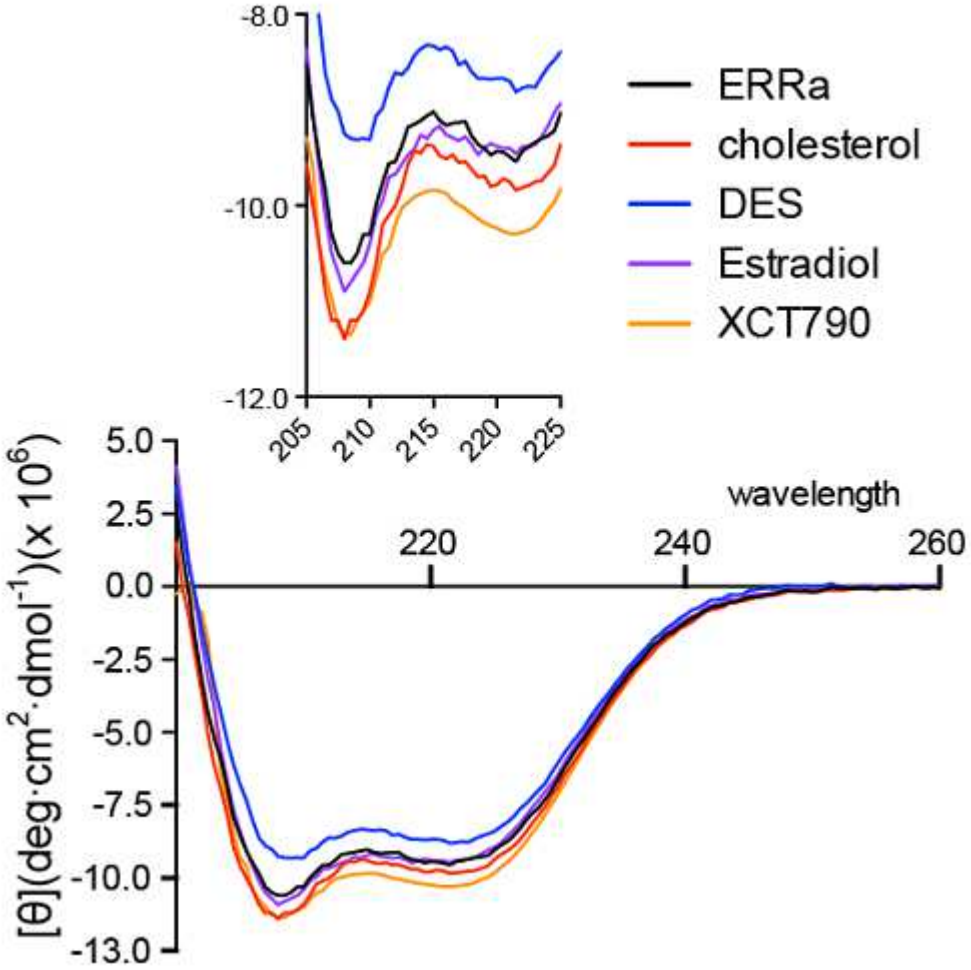


Figure 2.4: Cholesterol alters ERRα-LBD conformation. Cholesterol, and synthetic ERRα antagonists alter the conformation of ERRα-LBD as measured by circular dichroism. Estradiol, a ligand known not to bind ERRα, does not alter its conformation.

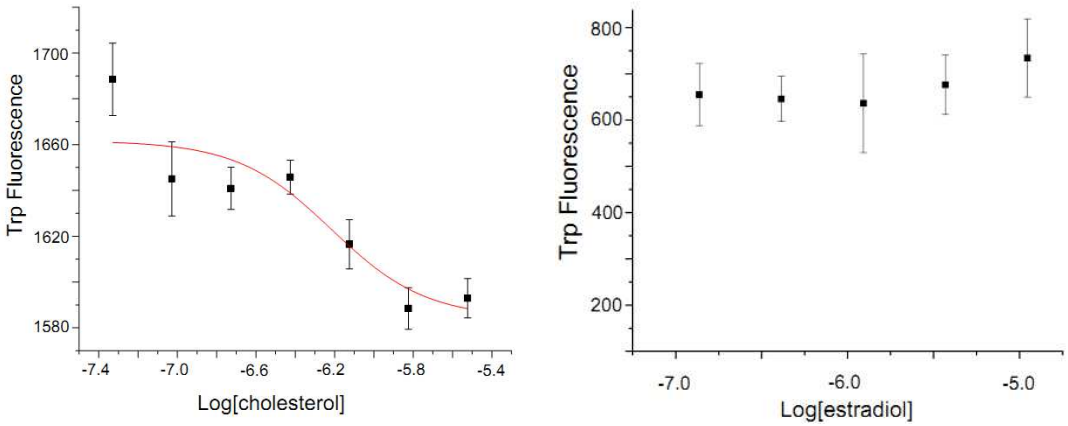


Figure 2.5: Cholesterol binds ERRα with high affinity. (Left) Intrinsic tryptophan fluorescence of ERRalpha upon addition of cholesterol demonstrated cholesterol bound with a kD of ~700nM. (Right) Intrinsic tryptophan fluorescence of ERRα does not change upon addition of a negative control sterol, estradiol.

2.3 Structural Analysis of ERRα Cholesterol Binding

To gain additional detail into how cholesterol binds ERRα, we probed ERRα binding with two different fluorescent cholesterol derivatives, 25-NDB- and 6-NDB-cholesterol. While 25-NDB-cholesterol binds HIS-ERRα-LBD, 6-NDB-cholesterol did not (Fig. 2.6). The data suggest that cholesterol is oriented with its hydroxyl group facing into the ligand-binding pocket. We obtained a more detailed structural model for cholesterol binding by computationally docking cholesterol into the ERRα ligand-binding pocket. These simulations identified a hydrogen bond between amino acid E331 and the cholesterol hydroxyl group (Fig 2.6). In addition, F328 and L324 also appear to make important hydrophobic contacts with cholesterol. Mutagenesis studies supported this model. Mutating E331A, F328A, and L324A left the secondary structure of ERRα intact, but prevented cholesterol binding (Fig 2.6).

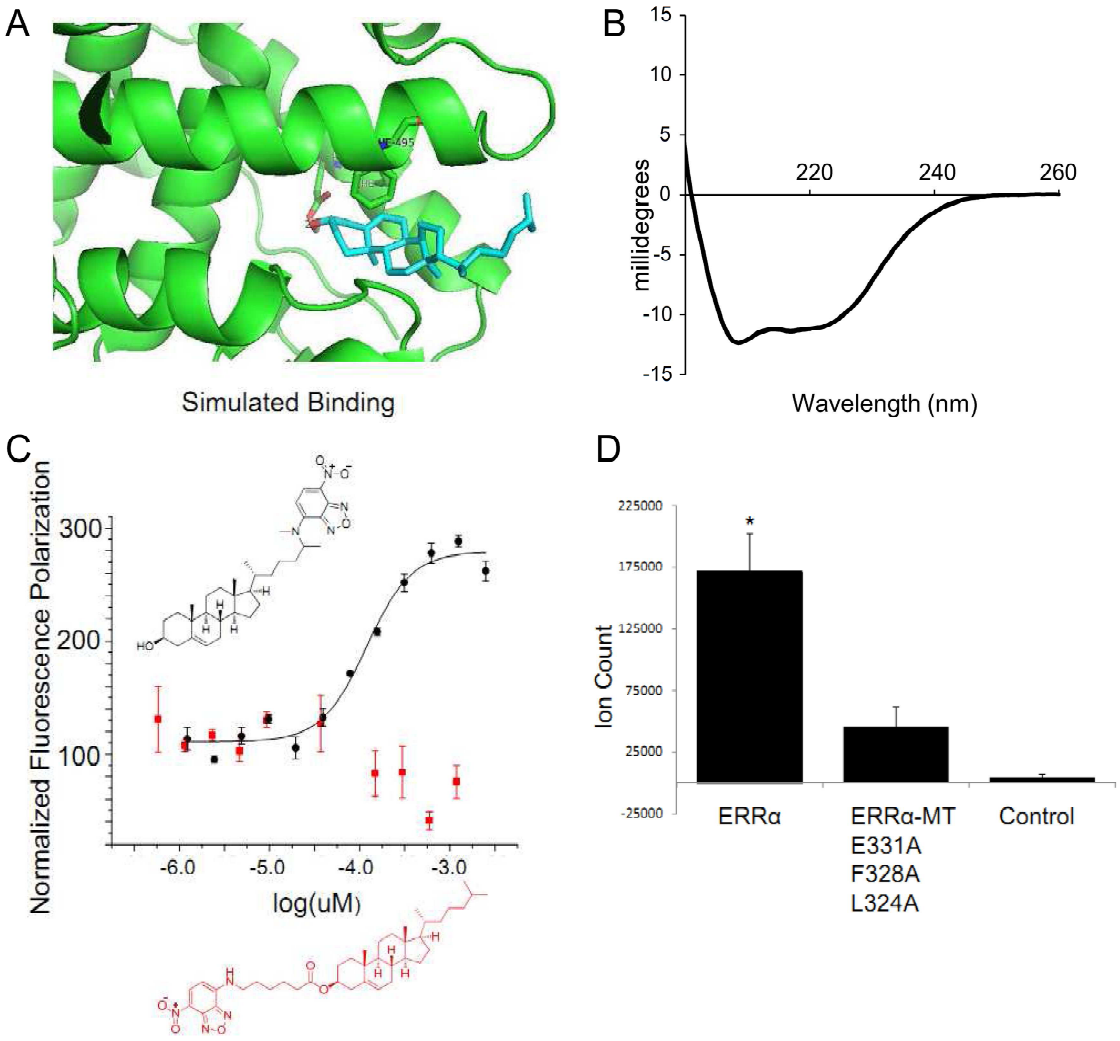


Figure 2.6: Cholesterol binds in the ligand binding pocket of ERRα. (A) Docking of cholesterol to the ERRα ligand binding pocket suggests cholesterol has important interactions with three amino acids in the ERRα ligand binding pocket, E331, F328, and L324. (B) An ERRα triple mutant E331A, F328A, and L324A has alpha helical secondary structure indicative of proper folding as indicated by circular dichroism. (C) Fluorescence polarization assays demonstrated that cholesterol fluorescently labeled at the aliphatic chain, but not at the sterol are able to bind to ERRα indicating that cholesterol binds with its hydroxyl end facing into the ligand binding pocket. (D) The ERRα E331A F328A L324A triple mutant can no longer bind cholesterol relative to the WT protein. Statistics was performed with a Student's t-test, *, $p \leq 0.05$.

2.4 Cholesterol regulates ERRα Transcription

If cholesterol is a bona fide ERRα ligand, it must demonstrate the ability to modulate ERRα transcriptional activity. Conventional luciferase transactivation strategies, where a ligand is added to detect an increase in transcriptional activity, are complicated by the presence of abundant amounts of endogenous ligands, such as cholesterol, already in the media and cells. Rather than introduce cholesterol, we decided that a more prudent strategy would be to measure ERRα activity upon cholesterol depletion. We lowered intracellular cholesterol levels in three ways: lipid free serum, cholesterol-binding cyclodextrans and/or statins. Cyclodextran and statin both attenuated ERRα transactivation, and also had a cumulative effect (Fig. 2.7). This data indicate that cholesterol levels positively correlate with ERRα activity, which would be expected for an ERRα agonist. Importantly, adding back of cholesterol to these samples rescued ERRα activity, indicating that cholesterol and not a precursor of the cholesterol pathway is involved in ERRα activity (Figure 2.7). These results suggest that cholesterol is an endogenous ERRα agonist that promotes ERRα-mediated transcription.

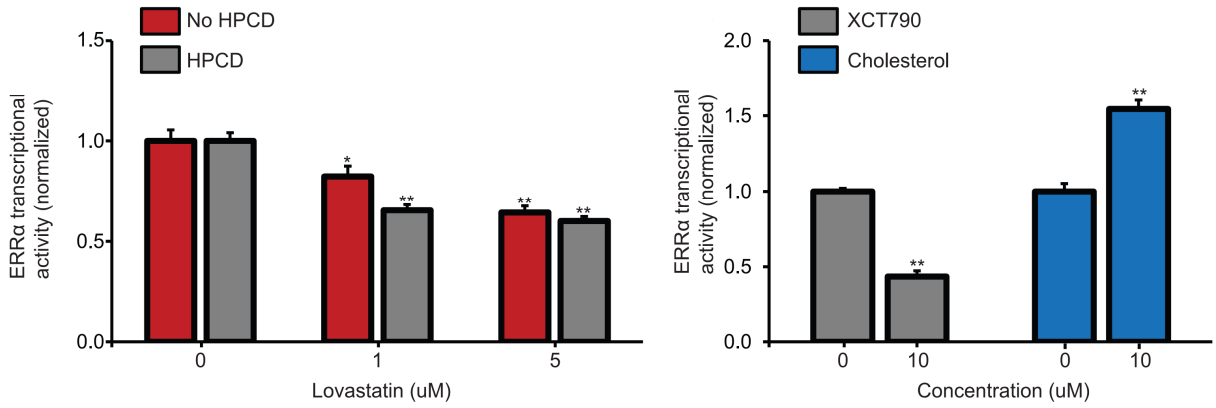


Figure 2.7: Cholesterol is an ERRα agonist. Luciferase assays demonstrate that upon depletion of cholesterol with lovastatin or HPCD ERRα transactivation is attenuated. Addition of cholesterol to cholesterol depleted cells rescues ERRα transactivation. Statistics was performed with a Student's t-test, *, p ≤ 0.05, **, p≤ 0.01.

2.5 Cholesterol regulates ERRα activity in Osteoclastogenesis

On the basis of these results, we next investigated whether cholesterol and ERRα are functionally dependent. We decided to examine osteoclastogenesis, a key cellular differentiation process in the physiological regulation of bone remodeling as well as in diseases such as osteoporosis. Osteoclasts are differentiated from monocyte/macrophage precursor cells upon treatment with receptor activator of nuclear factor kappa-B ligand (RANKL)[17,18]. We have recently shown that ERRα promotes osteoclast differentiation and activity, as a result, ERRα knockout mice exhibit decreased bone resorption and high bone mass[19]. Interestingly, both statins and nitrogen-containing bisphosphonates suppress osteoclast function; and both inhibit the cholesterol synthesis pathways, by blocking the HMG-CoA reductase and farnesyl

diphosphate synthase (FPPS), respectively[20,21]. However, the molecular target for how exactly the cholesterol synthesis pathway impacts osteoclastogenesis is still an open question, although interference with protein prenylation by bisphosphonates has been suggested[21].

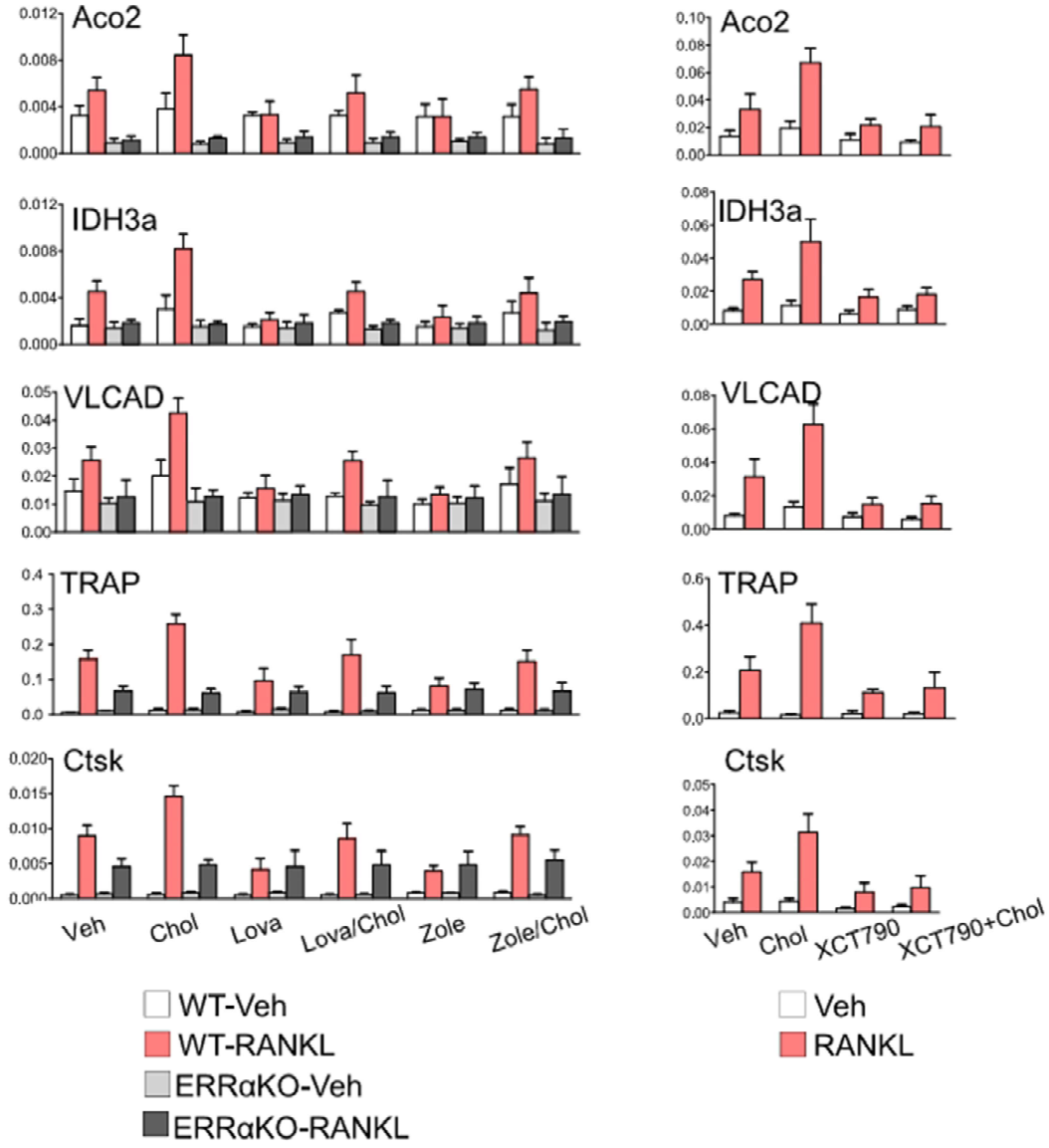


Figure 2.8: Cholesterol regulates osteoclastogenesis markers through ERRα. (Left) Cholesterol addition or depletion regulates osteoclast marker genes in the presence but not absence of ERRα. (Right) WhenERRα cholesterol binding is blocked by addition of a synthetic ERRα antagonist, XCT790, cholesterol does not affect osteoclast marker gene levels.

To test the hypothesis that cholesterol, statins and bisphosphonates regulate osteoclast differentiation via ERRα, we performed bone marrow osteoclast differentiation assays. We found that osteoclastogenesis was strongly dependent on the presence or absence of cholesterol. Cholesterol addition enhanced osteoclast formation, whereas cholesterol depletion by treatment with zoledronate (a clinically used nitrogen-containing bisphosphonates) or lovastatin inhibited osteoclast formation (Figure 2.8). Cholesterol adding back to zoledronate- or lovastatin-treated cells was able to rescue osteoclast differentiation indicating cholesterol rather than cholesterol precursors in the synthesis pathway was responsible for the changes in differentiation (Figure 2.8). Gene expression analysis showed that cholesterol up-regulated the expression of both ERRα target genes that stimulate mitochondrial biogenesis and osteoclast activity, such as Aco2, IDH3a and VLCAD[3], and osteoclast differentiation markers such as TRAP (tartrate-resistant acid phosphatase) and Ctsk (Cathepsin K) (Figure 2.8).

Similar experiments using ERRα null (ERRα -/-) bone marrow osteoclast differentiation cultures reveal that the effects of cholesterol on osteoclastogenesis was completely abolished by ERRα deletion. In the absence of ERRα, osteoclast differentiation was neither enhanced by cholesterol nor suppressed by lovastatin or zoledronate (Fig. 3a). Consistent with this observation, cholesterol, lovastatin or zoledronate no longer altered the expression of ERRα target genes or osteoclast differentiation markers (Fig. 3b). In addition to genetic loss-of-function by ERRα deletion, biochemical inhibition with a synthetic ERRα antagonist/inverse agonist XCT790 also prevented the effects of cholesterol on WT osteoclast differentiation

cultures (Fig. 3c). These data indicate that cholesterol enhances osteoclastogenesis in an ERRα-dependent manner. In conjunction with biochemical data demonstrating cholesterol binding to ERRα-LBD and cellular reporter assays demonstrating cholesterol stimulation of ERRα transcriptional activity, these results suggest that cholesterol promotes osteoclastogenesis by functioning as an ERRα agonist, whereas statins and bisphosphonates suppress osteoclastogenesis by reducing the bioavailability of the endogenous ERRα agonist cholesterol.

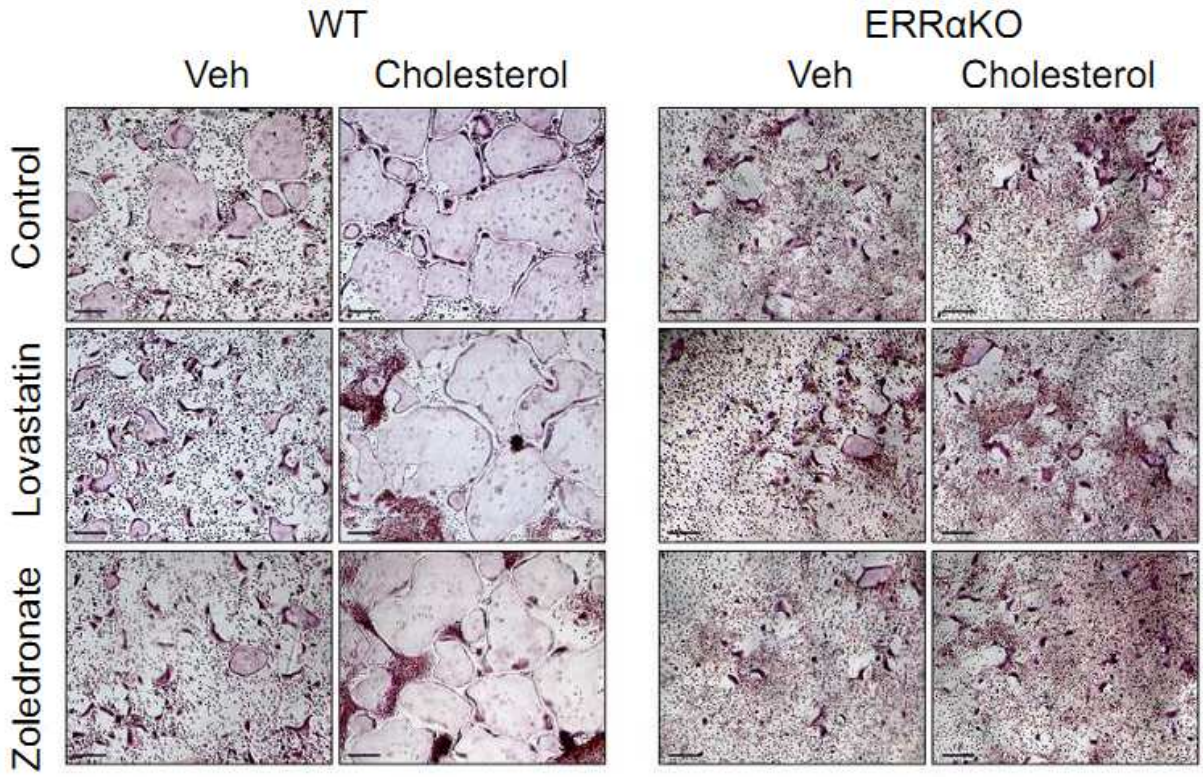


Figure 2.9 Cholesterol regulates osteoclastogenesis in the presence of ERRα: Cholesterol addition or depletion with statins or bisphosphonates regulated osteoclastogenesis in macrophages (Left), but not in ERRα knockout macrophages (right). Osteoclasts are the large pink cells. Undifferentiated macrophages are the smaller pink dots.

2.6 Cholesterol agonism reveals a novel role for ERRα in atherosclerotic foam cell formation

We next investigated whether the recently reported anti-inflammatory role of cholesterol in macrophages was also ERRα-dependent, to further examine the functional association between ERRα and cholesterol in a different biological context. Cholesterol has been shown to inhibit macrophage expression of cytokines Cxcl9 and Cxcl10[22]. We found that LPS-induced Cxcl9 and Cxcl10 expression was suppressed by cholesterol and exacerbated by lovastatin or zoledronate in WT macrophages. In contrast, these effects were once again abolished in ERRα-/- macrophages (Figure 2.10). Furthermore, cholesterol inhibition of Cxcl9 and Cxcl10 expression was also absent in macrophages treated with XCT790. These data not only indicate that the cytokine-suppressive effect of cholesterol in macrophages is ERRα-mediated but also reveal a novel anti-inflammatory role for ERRα. Importantly, these results provide further evidence that cholesterol is a functional ERRα agonist in another biological process.

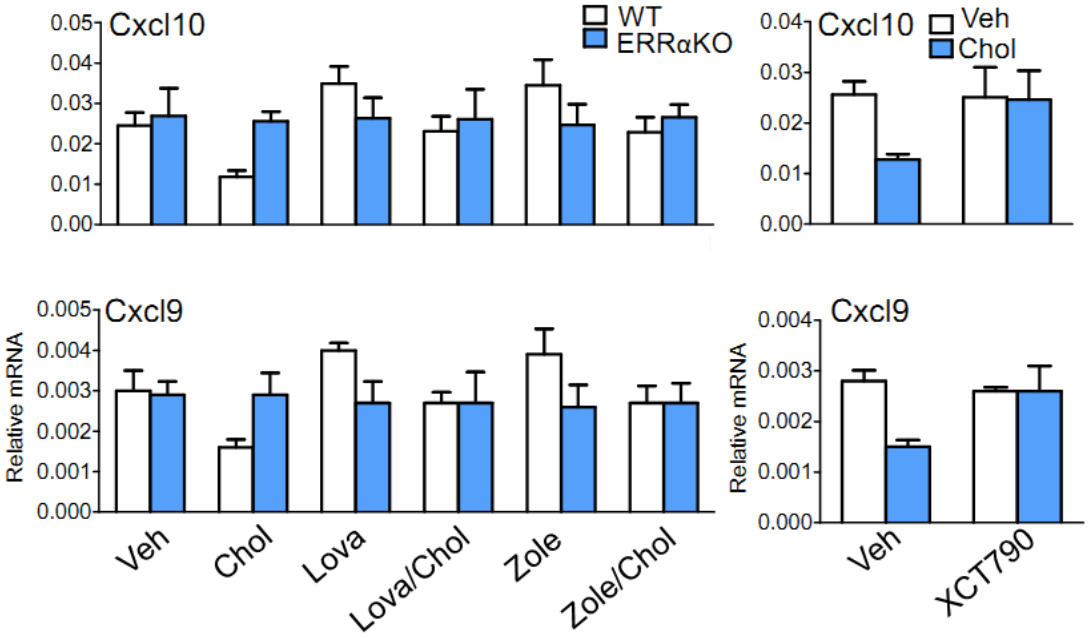


Figure 2.10 Cholesterol regulates inflammation in an ERRα dependent manner. Addition or depletion regulates mRNA levels of markers involved atherosclertoic inflammation in macrophages. However, this effect is dependent on ERRα, and cholesterol does not affect inflammation in the absence of ERRα. Likewise, the effects of cholesterol are blocked when cholesterol ERRα binding is blocked by addition of a synthetic antagonist ERRα.

2.7 Cholesterol functions as an ERRα agonist *in vivo*

To gain insight to whether the functional relationship between cholesterol and ERRα extends to physiology and pharmacology, we next performed *in vivo* studies in the context of bone resorption and skeletal remodeling. Epidemiological and pharmacological studies in human show that cholesterol-lowering drugs such as bisphosphonates and statins are bone protective[14], whereas hypercholesterolemia is associated with bone loss[20,21]. As a loss-of-cholesterol-function approach, we treated WT or ERRα-/- mice with zoledronate, which targets bone more efficiently than statins, and compared them to vehicle-treated mice 4 weeks later. ELISA analyses showed that zoledronate significantly

decreased the serum levels of the bone resorption marker C-terminal telopeptide fragments of the type I collagen (CTX-1) in WT mice; in contrast, this effect was completely abolished in the ERRα-/- mice (Figure 2.11a). Moreover, CTX-1 was reduced to a similar extent by both cholesterol ligand deletion in the zoledronate-treated mice (-82%) and by receptor deletion in the ERRα-/- mice (-80%) (Figure 2.11a). In line with these results, microCT analysis of the proximal tibiae show that zoledronate significantly increased bone mass in WT mice but not in ERRα-/- mice (Figure 2.11b-c). Furthermore, bone histomorphometry analysis show that zoledronate significantly reduced osteoclast surface and osteoclast number in WT mice but not in ERRα-/- mice (Figure 2.11d).

Complementarily, as a gain-of-cholesterol-function approach, we fed WT or ERRα-/- mice with a high-cholesterol-diet (HCD) for 4 weeks and compared them to chow-diet fed control mice. ELISA analyses showed that HCD feeding significantly elevated serum CTX-1 levels by 107% in WT mice, once again, this effect was completely abolished in the ERR-/- mice (Figure 2.11e). In agreement, microCT analysis of the proximal tibiae show that HCD feeding significantly decreased bone mass in WT mice but not in ERRα-/- mice (Figure 2.11f-g). Furthermore, bone histomorphometry analysis show that HCD feeding significantly increased osteoclast surface and osteoclast number in WT mice but not in ERRα-/- mice (Figure 2.11h). These *in vivo* studies indicate that the physiological function of cholesterol and the pharmacological actions of cholesterol-lowering drugs requires ERRα, further supporting the notion that cholesterol is a functional endogenous ERRa agonist.

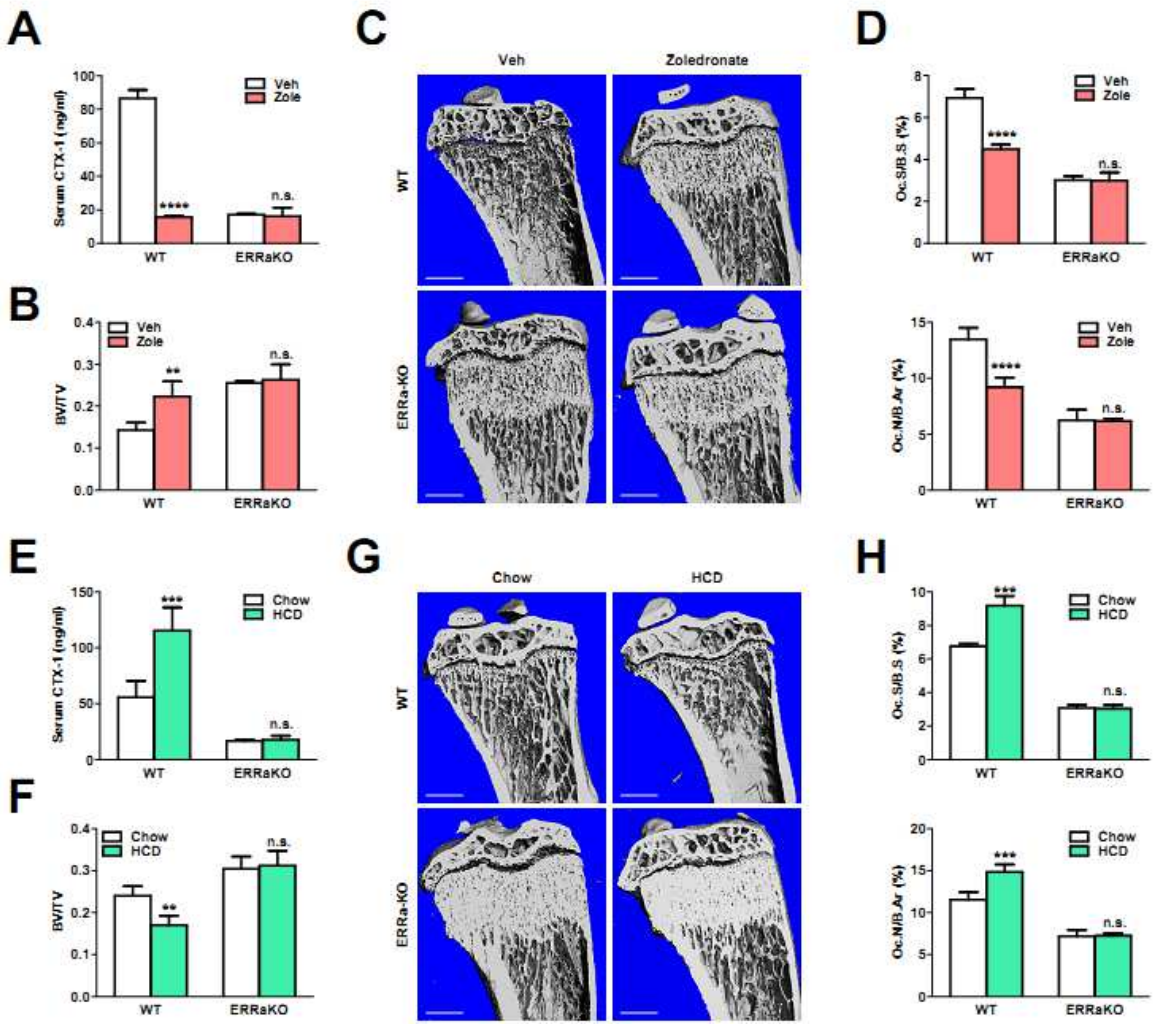


Figure 2.11: ERRα mediates cholesterol function in vivo. Osteoprotective effects of zoledronate was abolished in ERRα-/- mice. WT or ERRα-/- mice (6 week old male, n=4) were treated with a single i.v. injection of zoledronate at 0.54 mg/kg or PBS vehicle control and analyzed 4 weeks later. **a,** Serum levels of the bone resorption marker CTX-1. **b-c,** MicroCT analysis of tibiae. **b,** Trabecular bone volume/tissue volume ratio (BV/TV). **c,** Representative CT images of the entire proximal tibia (scale bar, 1mm). **d,** Bone histomorphometry analysis of osteoclast surface/bone surface (Oc.S/BS) (top) and osteoclast number/bone area (Oc.N/B.Ar) (bottom). **e-h,** Bone loss induced by high cholesterol diet (HCD) feeding was abolished in ERRα-/- mice. WT or ERRα-/- mice (7 week old female, n=4) were fed with a HCD or chow control diet for 4 weeks. **e,** Serum levels of the bone resorption marker CTX-1. **f-g,** MicroCT analysis of tibiae. **f,** Trabecular BV/TV. **g,** Representative CT images of the entire proximal tibia (scale bar, 1mm). **h,** Bone histomorphometry analysis of Oc.S/BS (top) and Oc.N/B.Ar (bottom). Statistical analyses were performed with Student's t-Test and are shown as mean ± standard deviation; **, $p<0.01$; ***, $p<0.005$; ****, $p<0.001$; n.s., non-significant ($p>0.05$).

2.8 Conclusion

In conclusion, our metabolomics strategy has successfully deorphanized ERRα by identifying cholesterol as an endogenous ERRα ligand. This discovery led to the elucidation of a key mechanistic link for how cholesterol and ERRα regulate osteoclastogenesis: cholesterol promotes osteoclastogenesis by activating its receptor ERRα; and ERRα stimulates osteoclastogenesis by responding to a rise in cellular cholesterol level. Moreover, our findings illuminate new mechanistic insights for how clinically used drugs such as statins and bisphosphonates actually act. Particularly, our results question the long-standing hypothesis that the inhibition of osteoclast activity by bisphosphonates is due solely to the blockade of protein prenylation. Interestingly, there is crosstalk between ERRα and ERα, including recognition of the estrogen response element (ERE) by ERRα. Thus, in ER negative (*ER*-) breast cancer, the presence of ERRα is a negative prognostic factor as it compensates for the loss of ERα in addition to triggering the expression of ERα-independent genes[23]. The function of cholesterol as an ERRα ligand may also provide mechanistic insight into clinical studies suggesting that statins can be used to treat or prevent *ER*- breast cancer[24,25]. The presence or absence of ERRα in *ER*- breast cancer may be a prognostic factor for the effectiveness of cholesterol-lowering treatments as anti-cancer therapeutics. Although it remains to be seen whether modulation of ERRα transactivation through therapeutic regulation of cholesterol is universally effective in the clinical setting, the deorphanization of ERRα using a widely applicable approach opens a new frontier in nuclear receptor biology.

2.9 Methods

Lipid Isolation:

To mouse brain or kidney tissue 2 mL chloroform, 1 mL water, and 1 mL methanol was added. Tissue was then dounced to homogeneity. Solvent was collected and mortar and pestle were washed with an additional 1 mL chloroform, .5 mL water, and .5 mL methanol. Fractions were pooled and centrifuged for 10 minutes at 3220 g. The chloroform layer was collected and dried under nitrogen. Lipids were then redissolved in DMSO at 40ul/mg.

Sterol Isolation:

Sterols were prepared in the same manner as total lipids were except after drying under nitrogen sterols were redissolved in 1 mL toluene. They were then fractionated from total lipids using solid phase extraction (BJ9050). The column was equilibrated with 1 mL hexanes, and then loaded with the lipid sample. The column was then washed with hexanes (1 mL) and sterols were eluted with 30% isopropanol in hexanes (8 mL). After elution the sample was dried under nitrogen and redissolved to 40uL/mg.

A 4% solution (v/v) of sterol mix in buffer (20mM Tris, 200mM NaCl, pH 8.0) was made, and 20uL: of this lipid sample was injected on LC-MS (Agilent 6140) to check that sterols were dissolved. In all cases cholesterol ion intensity was greater than 200,000 counts.

Recombinant Protein Expression

The HIS-ERRα-LBD plasmid reported by Greshik et al[26] was provided courtesy of Dino Moras. HIS-ERRα-LBD was expressed in Rosetta 2 (DE3) cells. Terrific broth was inoculated with a 1:50 dilution of starter culture. Cells were grown to OD .4 at 37°C and

then moved to 18°C. Once they reached OD .8 they were induced with .1mM IPTG and grown for 16 hours before harvesting.

Purification of HIS-ERRα-LBD:

HIS-ERRα-LBD was purified by nickel affinity chromatography and size exclusion chromatography.

Global lipid pull down:

Centrifuge columns (Pierce product #89868) were loaded with 10uL IMAC sepharose 6 Fast Flow beads (GE). Beads were washed with water (2 x 500uL) and the wash was eluted by centrifugation. 100uL .2M nickel sulfate was added to each column and the column was incubated for 30 minutes. Nickel sulfate was then eluted and columns were washed with water (3 x 100uL) and protein buffer (20mM Tris, 200mM NaCl, pH 8.0) (3 x 200uL). 200uL of 25uM or 50uM HIS-ERRα-LBD or protein buffer containing no protein was incubated with resin for 2 hours. After incubation, protein was eluted and beads were washed with protein buffer containing 50mM imidazole (3 x 200uL). 100uL buffer containing 4% or 12% lipid mix by volume was added to each column and incubated for 30 minutes. The lipid mixture was then eluted and columns were washed with protein buffer containing 50mM imidazole (3 x 100uL). Protein-lipid conjugate was then eluted with protein buffer containing 250mM imidazole (3 x 100uL). 80uL were injected on LC-MS TOF in negative or positive mode. This procedure was done in triplicate.

LC Gradient:

Solvent A: 95% Water, 5% Methanol

Solvent B: 60% Isopropanol, 35% Methanol, 5% water

Positive mode solvent modifier: 5mM ammonium formate .1% formic acid

Negative mode solvent modifier: .1% ammonium hydroxide.

A gradient from 0-100% solvent B was run over 58 minutes on a C4 biobond 5µm, 50 x 4.6 mm column.

Data Analysis:

After LC-MS TOF analysis, data from replicates was aligned using XCMS. Data was filtered for ions with greater than 100,000 average counts in either the protein or resin samples. Ions that were not present in the lipid sample alone were also removed. To identify ions of interest, data was further filtered for a p value of less than .05 (1 tailed t-test) and a fold change of greater than 2.

Pull down for the targeted identification of sterols:

The ERRα sterol pulldown was performed as the global lipid pulldown above except a sterol mixture was used in place of a lipid mixture. 4% (v/v) sterol mixture was used in 20mM Tris, 200mM NaCl pH 8.0 buffer for the sterol pulldown. After elution, aliquots were extracted into chloroform using a chloroform-methanol-water mixture (2:1:1) and dried down under nitrogen. Then they were redissolved in 30uL water and 20uL was injected on LC-MS QQQ (Agilent 6140).

MRM sterol analysis method:

Our method to identify sterols was developed from work by Shan et al. and Mcdonald et al[27,28]. Sterols were analyzed on an Agilent 6140 LC-MS QQQ. Solvent A was 85% methanol, 15% water, and 5mM ammonium acetate. Solvent B was 100% methanol and 5mM ammonium acetate. A phenomenex C18 100Å 250x2mm 3 micron column was used.

This method to identify sterols was developed from work by Shan et al. and Mcdonald et al[27,28].

Table 2.1 MRM sterol analysis global profiling gradient.

Time	%B	Flow	Max pressure
0	0	0.25	500
12	0	0.25	500
45	100	0.25	500
65	100	0.25	500
65.1	0	0.25	500
78	0	0.25	500

MS/MS analysis:

Gas Temp: 100°C

Gas flow (l/min): 8

Nebulizer (psi): 35

Capillary (V) 4000

Table 2.2: MRM mass spectrometry method for the targeted identification of sterols.

Compound	Precursor ion	MS1 Res	Product ion	MS2 Res	Dwell	Frag (V)	CE (V)
Sat-tetraol	454.4	Wide	367.3	Unit	80	65	
Unsat-tetraol	452.4	Wide	365.3	Unit	80	65	
Lanosterol	444.4	Wide	409.4	Unit	80	95	
Sat-triol 1	438.4	Wide	420.4	Unit	80	65	
Sat-triol 2	438.4	Wide	385.3	Unit	80	65	
Sat-triol 3	438.4	Wide	367.3	Unit	80	65	
Unsat-triol 1	436.4	Wide	418.4	Unit	80	65	
Unsat-triol 2	436.4	Wide	383.3	Unit	80	65	
Unsat-triol 3	436.4	Wide	365.3	Unit	80	65	
24-dihydrolanosterol	429.4	Wide	411.4	Unit	80	115	
hydroxy-epoxy mix 1	420.4	Wide	385.3	Unit	80	65	
hydroxy epoxy mix 2	420.4	Wide	367.3	Unit	80	65	
24-25 epoxychol	418.4	Wide	383.3	Unit	80	60	
cholestanol	406.4	Wide	371.4	Unit	80	125	
lathosterol/cholesterol	404.4	Wide	369.4	Unit	80	125	
desmosterol	402.4	Wide	367.3	Unit	80	70	
7-ketochol	401.3	Wide	383.3	Unit	80	95	
unsat-mix	385.3	Wide	367.3	Unit	80	100	
7alpha OH cholesterol	385.3	Wide	367.3	Unit	80	65	

Diethylstilbestrol enrichment:

The lipid mixture was spiked with 79uM diethylstilbestrol (DES), and the global lipid pulldown procedure above was repeated with this mixture. After LC-MS TOF, diethylstilbestrol levels were analyzed in the extracted ion chromatogram.

Diethylstilbestrol inhibition of cholesterol binding:

The sterol mixture was spiked with 79uM DES and the pull down for the targeted identification of sterols was repeated as above using this mixture.

Circular Dichroism:

Solutions of 6.36uM HIS-ERRα-LBD in 20mM Tris, 200mM NaCl pH 8.0 2% ethanol (v/v) were incubated with 100uM small molecule for 5 minutes. Then circular dichroism

was measured on JASCO J-710 spectrophotometer. Four measurements were averaged.

Tryptophan Fluorescence Quenching:

Cholesterol was titrated into 350uL of .6uM HIS-ERRα-LBD. The sample was excited at 280nm and fluorescence emission was collected at 333nm with a cutoff filter at 325nm.

Docking Simulations:

Docking simulations were conducted with autodock vina (vina.scripps.edu) using the ERRα LBD crystal structure 2JPL in the protein data bank. A cholesterol structure was generated in ChemBioDraw. The simulation was centered at x coordinate -12, y coordinate -10, z coordinate 12. Simulation box size was x=22, y=22, z=22. The simulation was performed with a step size (exhaustiveness) of 1000.

Fluorescence Polarization Experiments:

25-NDB cholesterol and 6-NDB cholesterol were used at concentrations of 4.4uM and 3.8uM respectively in 20mM Tris 200mM NaCl pH 8.0 5% ethanol (v/v). HIS-ERRα-LBD was added and samples were excited at 480nm and emission was measured at 560nm for 25-NBD cholesterol and 550nm for 6-NBD cholesterol.

ERRα AAA sterol pulldown:

This lipid pulldown was performed as described above except sterols were dissolved in ethanol instead of DMSO.

Luciferase, and osteoclast differentiation, and knockout mice assays:

Transfection and luciferase reporter assay were performed in CV-1 cells as described[3]. Bone marrow osteoclast differentiation assays and bone analyses (microCT, ELISA and histomorphometry) were performed as described[3]. ERRα knockout mice on a C57B6/J background was previously described[3,4]. Zoledronate was administered by a single intravenous injection at 0.54 mg/kg. High cholesterol diet (2% cholesterol, 0.5% NaCholate) and normal chow diet control were from Harlan Teklad. All statistical analyses were performed with Student's t-Test and presented as mean ± standard deviation (s.d.). The p values were designated as: *, $p<0.05$; **, $p<0.01$; ***, $p<0.005$; ****, $p<0.001$; n.s. non-significant ($p>0.05$).

.

2.10 References

(1) Giguere, V. *Trends Endocrinol Metab* **2002**, *13*, 220.

(2) Giguere, V.; Yang, N.; Segui, P.; Evans, R. M. *Nature* **1988**, *331*, 91.

(3) Wei, W.; Wang, X.; Yang, M.; Smith, L. C.; Dechow, P. C.; Sonoda, J.; Evans, R. M.; Wan, Y. *Cell metabolism* **2010**, *11*, 503.

(4) Luo, J.; Sladek, R.; Carrier, J.; Bader, J. A.; Richard, D.; Giguere, V. *Mol Cell Biol* **2003**, *23*, 7947.

(5) Patch, R. J.; Searle, L. L.; Kim, A. J.; De, D.; Zhu, X.; Askari, H. B.; O'Neill, J. C.; Abad, M. C.; Rentzeperis, D.; Liu, J.; Kemmerer, M.; Lin, L.; Kasturi, J.; Geisler, J. G.; Lenhard, J. M.; Player, M. R.; Gaul, M. D. *J Med Chem* **2011**, *54*, 788.

(6) Sladek, R.; Bader, J. A.; Giguere, V. *Mol Cell Biol* **1997**, *17*, 5400.

(7) Giguere, V. *Endocr Rev* **2008**, *29*, 677.

(8) Ascenzi, P.; Bocedi, A.; Marino, M. *Mol Aspects Med* **2006**, *27*, 299.

(9) Horard, B.; Vanacker, J. M. *J Mol Endocrinol* **2003**, *31*, 349.

(10) Kallen, J.; Schlaeppi, J. M.; Bitsch, F.; Filipuzzi, I.; Schilb, A.; Riou, V.; Graham, A.; Strauss, A.; Geiser, M.; Fournier, B. *J Biol Chem* **2004**, *279*, 49330.

(11) Wu, F.; Wang, J.; Wang, Y.; Kwok, T. T.; Kong, S. K.; Wong, C. *Chem Biol Interact* **2009**, *181*, 236.

(12) Duellman, S. J.; Calaoagan, J. M.; Sato, B. G.; Fine, R.; Klebansky, B.; Chao, W. R.; Hobbs, P.; Collins, N.; Sambucetti, L.; Laderoute, K. R. *Biochem Pharmacol* **2010**, *80*, 819.

(13) Lee, J. M.; Lee, Y. K.; Mamrosh, J. L.; Busby, S. A.; Griffin, P. R.; Pathak, M. C.; Ortlund, E. A.; Moore, D. D. *Nature* **2011**, *474*, 506.

(14) Ayers, S. D.; Nedrow, K. L.; Gillilan, R. E.; Noy, N. *Biochemistry* **2007**, *46*, 6744.

(15) Chakravarthy, M. V.; Lodhi, I. J.; Yin, L.; Malapaka, R. R.; Xu, H. E.; Turk, J.; Semenkovich, C. F. *Cell* **2009**, *138*, 476.

(16) Kim, Y. G.; Lou, A. C.; Saghatelian, A. *Mol Biosyst* **2011**, *7*, 1046.

(17) Boyle, W. J.; Simonet, W. S.; Lacey, D. L. *Nature* **2003**, *423*, 337.

(18) Novack, D. V.; Teitelbaum, S. L. *Annu Rev Pathol* **2008**, *3*, 457.

(19) Ayers, S. D. N., Katherine L.; Gillilan, Richard E.; Noy, Noa *Biochemistry* **2007**, *46*, 6744.

(20) Toledano, J. E.; Partridge, N. C. *Trends Endocrinol Metab* **2000**, *11*, 255.

(21) Russell, R. G. *Bone* **2011**, *49*, 2.

(22) Spann, N. J.; Garmire, L. X.; McDonald, J. G.; Myers, D. S.; Milne, S. B.; Shibata, N.; Reichart, D.; Fox, J. N.; Shaked, I.; Heudobler, D.; Raetz, C. R.; Wang, E. W.; Kelly, S. L.; Sullards, M. C.; Murphy, R. C.; Merrill, A. H., Jr.; Brown, H. A.; Dennis, E. A.; Li, A. C.; Ley, K.; Tsimikas, S.; Fahy, E.; Subramaniam, S.; Quehenberger, O.; Russell, D. W.; Glass, C. K. *Cell* **2012**, *151*, 138.

(23) Suzuki, T.; Miki, Y.; Moriya, T.; Shimada, N.; Ishida, T.; Hirakawa, H.; Ohuchi, N.; Sasano, H. *Cancer Res* **2004**, *64*, 4670.

(24) Cuzick, J.; DeCensi, A.; Arun, B.; Brown, P. H.; Castiglione, M.; Dunn, B.; Forbes, J. F.; Glaus, A.; Howell, A.; von Minckwitz, G.; Vogel, V.; Zwierzina, H. *Lancet Oncol* **2011**, *12*, 496.

(25) Kumar, A. S.; Benz, C. C.; Shim, V.; Minami, C. A.; Moore, D. H.; Esserman, L. J. *Cancer Epidemiol Biomarkers Prev* **2008**, *17*, 1028.

(26) Greschik, H.; Althage, M.; Flaig, R.; Sato, Y.; Chavant, V.; Peluso-Iltis, C.; Choulier, L.; Cronet, P.; Rochel, N.; Schüle, R.; Strömstedt, P.-E.; Moras, D. *Journal of Biological Chemistry* **2008**, *283*, 20220.

(27) Shan, H.; Pang, J.; Li, S.; Chiang, T. B.; Wilson, W. K.; Schroepfer Jr., G. J. *Steroids* **2003**, *68*, 221.

(28) McDonald, J. G.; Thompson, B. M.; McCrum, E. C.; Russell, D. W. *Methods in Enzymology* **2007**, *432*, 145.

Chapter 3: Discovery and Characterization of sORF Encoded Peptides

This chapter was adapted from:

Slavoff, SA.; Mitchell, AJ.*; Schwaid, AG.*; Cabili, MN; Ma, J.; Levin, JZ; Karger, AD.; Budnik, BA.; Rinn, JL; Saghatelian, A. Peptidomic Discovery of Short Open Reading Frame-Encoded Peptides in Human Cells. *Nature Chemical Biology* **2013** 9(1), 59.

*authors contributed equally

3.1 Introduction

The complexity of the small proteome remains incompletely explored because genome annotation methods generally break down for small open reading frames (ORFs), generally with a length cutoff of 100 amino acids. Computational[1] and ribosome profiling[2] studies have suggested that thousands of these non-annotated mammalian sORFs are translated. However, since these studies did not directly detect the presence of any sORF-encoded polypeptides (SEPs), it remains unknown whether sORFs produce polypeptides that persist in cells at biologically relevant concentrations, or are rapidly degraded. Indeed, biochemical analysis of the translation of two sORFs identified in the yeast GCN4 gene by ribosome profiling revealed that only one expressed detectable polypeptide product[3]. Moreover, recent evidence indicates that ribosome profiling can lead to widespread false positive identification of sORFs that encode polypeptides, and that mass spectrometry is the only surefire method to identify SEPs[4].

If SEPs do exist at physiologically relevant concentrations in cells, they may execute biological functions. Short open reading frames (sORFs) in the 5'-untranslated region (5'-UTR) of eukaryotic mRNAs (uORFs) are well studied[5-7] and some have been shown to produce detectable polypeptides[8,9]. In addition to uORFs, other sORFs in bacteria[10], viruses[11], plants[12,13], *Saccharomyces cerevisiae*[14], *Caenorhabditis elegans*[15], insects[16,17], and humans[18] have recently been discovered to produce polypeptides. Notably, the peptides encoded by the polycistronic *tarsel-less* (*tal*) gene in *Drosophila*, which are as short as 11 amino acids, regulate fly morphogenesis[16,17].

While no general method for discovering SEPs exists, attempts have been made to systematically identify these molecules. In *E. coli*, for example, experiments in which predicted sORFs were epitope-tagged revealed 18 SEPs[19]. In another example, a combination of computational and experimental approaches identified 299 potentially coding sORFs in *S. cerevisiae*, four of which were confirmed to produce protein and 22 of which appeared to regulate growth[14]. In human cells, an unbiased proteomics approach identified a total of four SEPs (defined here as polypeptides that are synthesized on the ribosome at a length of less than 150 amino acids) between the K562 and HEK293 cell lines with a length distribution of 88-148 amino acids[20]. The discordance between the small number of SEPs detected in human cells[20] and the large number of coding sORFs described by ribosome profiling[2] and computational methods[1] leaves open the possibility that SEPs are not produced as predicted or are rapidly degraded and therefore not detectable.

To resolve this question we developed of a novel SEP discovery and validation strategy that combines peptidomics and massively parallel RNA sequencing (RNA-Seq) (Figure 3.1). This strategy uncovered 90 SEPs, 86 novel SEPs, the largest number of human SEPs ever reported, which demonstrates that SEPs are much more abundant that previously reported. In addition, characterization of the encoding sORFs revealed interesting non-canonical translation events that give rise to SEPs, including bicistronic expression and the use of non-AUG start codons. One SEP, derived from the DEDD2 gene, localizes to mitochondria, which suggests that SEPs could generally have specific cellular localizations and functions. Together, these results highlight SEPs as an interesting class of polypeptides within the human proteome.

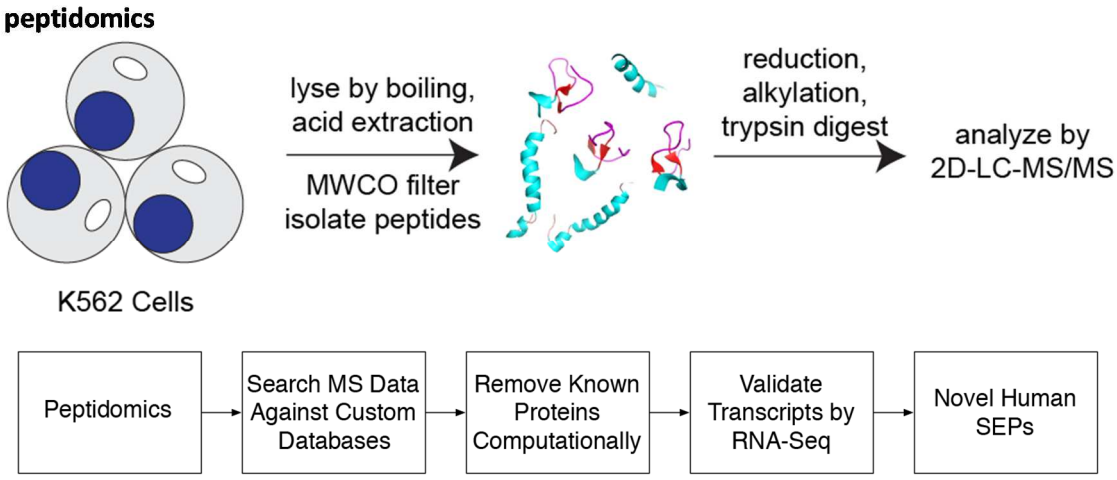


Figure 3.1 Workflow for identifying short ORF encoded peptides (SEPs).

3.2 Discovering SEPs Encoded by Annotated Transcripts

We developed a novel strategy that combines peptidomics and massively parallel RNA sequencing (RNA-Seq) to discover human SEPs (Figure 3.1). Peptidomics augments the traditional liquid chromatography-tandem mass spectrometry (LC-MS/MS) proteomics workflow to preserve and enrich small polypeptides[21]. In this context, the use of peptidomics increases the total number of SEPs detected, including a greater number of shorter SEPs. We isolated peptides from K562 cells, a human leukemia cell line, because we could use the previously reported SEPs in this cell line as positive controls[20]. Endogenous K562 polypeptides were isolated using our standard peptidomics workflow[21] with great care being taken to reduce proteolysis. Proteolysis is detrimental because the processing of cellular proteins greatly increases the complexity of the peptidome, which deteriorates the signal-to-noise ratio during the subsequent analysis[22]. After isolation, the K562 polypeptides were digested with trypsin and analyzed by LC-MS/MS (Figure 3.2). Based on previous results from our lab[23] and

others[24] the optimal size for detection by LC-MS/MS is approximately 10-20 amino acids, indicating that SEPs detection would greatly benefit from trypsin proteolysis.

To identify SEPs it was necessary to use a modified protocol for LC-MS/MS data analysis. Standard proteomics and peptidomics approaches identify peptides by matching experimentally observed spectra to databases of predicted spectra based on annotated genes, which would not include SEPs. We tore created a custom database containing all polypeptides that could possibly be translated from the human transcriptome (RefSeq). Using Sequest, an analysis program used to identify peptides from MS/MS spectra[25,26], we compared >200,000 MS/MS peptide spectra to this RefSeq-derived polypeptide database. This resulted in 6548 unique peptide identifications. We arrived at a tentative list of SEPs by keeping only those tryptic peptides that differed by at least two amino acids from every annotated protein to minimize the possibility of false positives arising from polymorphisms in annotated genes.

Due to the small size of SEPs, it is unlikely that an unbiased peptidomics experiment will detect more than one tryptic fragment of a given SEP, though eleven SEPs did have two or more fragments. This contrasts with standard proteomics studies, which, on account of the numerous tryptic fragments generated from full size polypeptides, will typically uncover two or more peptides to support the presence of a protein. Realizing that we would likely not be able to rely on the confidence contributed by the inherent redundancy of multiple-peptide protein identifications for SEP discovery, we submitted the candidate peptide spectrum matches (PSMs) to a rigorous evaluation procedure to ensure the highest confidence for each SEP.

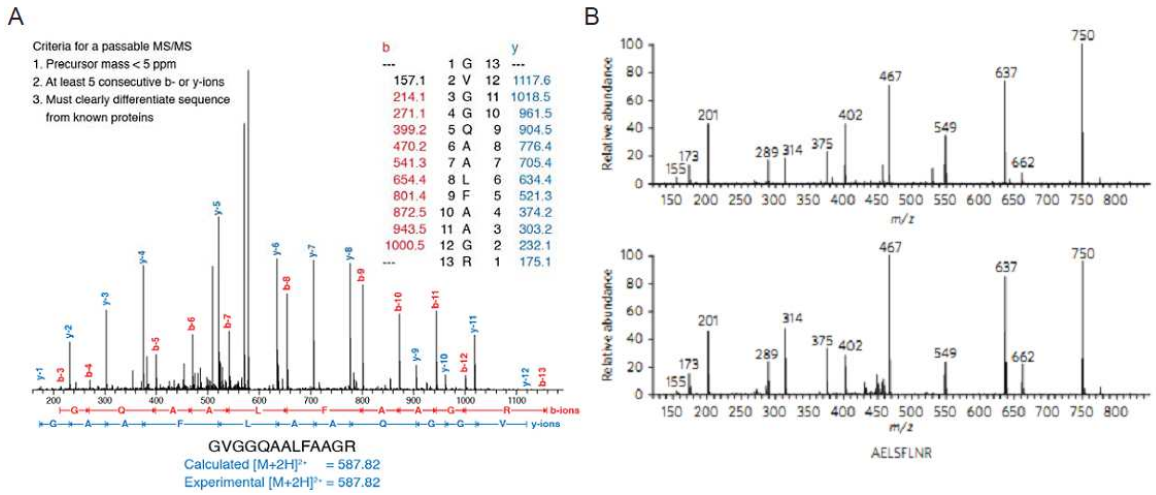


Figure 3.2: SEP MS/MS Criteria and Spectra. (A) Criteria for identifying SEPs. (B) We validated a select number of detected peptides by synthesizing them and comparing them to the endogenous spectra. (Top) Diagnostic spectra. (Bottom) Endogenous spectra.

First, we discarded any PSM with an Sf score of less than 0.75 (the threshold for a typical proteomics experiment is Sf < 0.4[27]). This eliminated over 95% of the candidate set. We then visually examined each remaining MS/MS spectrum to ensure that it met a stringent set of criteria (Figure 3.2). In particular, we required that there be a sequence tag of five consecutive b- or y-ions, a precursor mass error of <5 ppm, and sufficient sequence coverage to unambiguously differentiate each peptide from every annotated protein sequence. This step reduced the remaining peptide pool by approximately 75%, for a total of 39 putative SEPs. Our PSM evaluation procedure therefore selected the most confident ~1% of the peptide identifications in our original candidate set. As a check on the effectiveness of this procedure, we compared the experimentally-collected MS/MS spectra of several identified peptides to that of identical synthetic peptides (Figure 3.2).

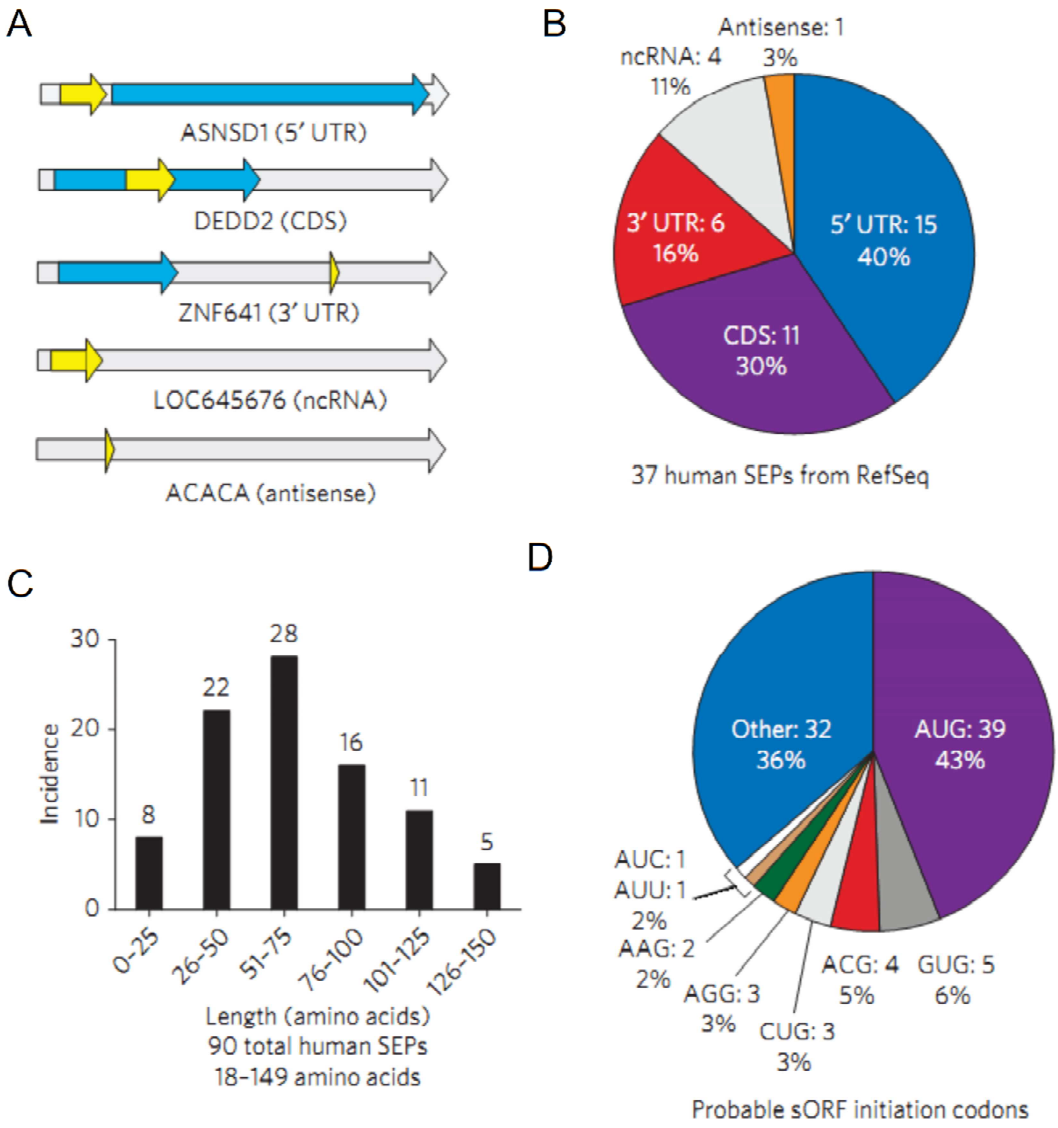


Figure 3.3: Overview of SEPs. (A) RNA maps illustrate the categories of sORFs that are translated into SEPs including 5' UTR, CDS, 3' UTR, ncRNA, and antisense sORFs. (B) Incidence of each category of SEPs within the RefSeq mRNAs. (C) Using protein databases derived from K562 RNA-Seq data revealed an additional 54 SEPs for a total of 90 human SEPs, 86 of which are previously uncharacterized. (D) Probable sORF initation codon usage. RNA maps are not to scale.

Lastly, to further reduce the probability of false positives, we comprehensively assembled and cataloged the K562 transcriptome using RNA-Seq and crosschecked the assembled RNA-Seq transcripts against our candidate sORF list. In this manner we

confirmed that at least 37 of the 39 implicated sORFs are present in this cell line and that no other sequence in the assembled K562 RNA-Seq transcripts could produce the detected peptides (Figure 3.3). This eliminated the possibility that the detected SEPs arose from point mutations in annotated genes, longer unannotated ORFs containing identical tryptic peptides, or post-transcriptional modification or editing of RNAs. Importantly, a similar analysis without trypsin failed to identify any SEPs demonstrating the importance of trypsin in generating an ideal sample for LC-MS/MS.

The 37 SEPs discovered through analysis of RefSeq transcripts fall into five major categories: (i) those located in the 5'-UTR, (ii) those located in the 3'-UTR, (iii) those located (frameshifted) inside the main coding sequence (CDS), (iv) those located on non-coding RNAs (ncRNAs), and (v) those located on antisense transcripts (Figure 3.3 A and B). The locations of these sORFs mirror the distribution obtained from ribosome profiling[2], indicating that our peptidomics coverage achieves the necessary breadth and depth to reveal global properties of sORFs (Figure. 3.3 B). Many of these SEPs appear to be derived from polycistronic mRNAs, which is interesting because this phenomenon has historically been thought to be rare in eukaryotes. However, our findings here are again consistent with those of ribosome profiling studies[2].

3.3 SEPs are Derived from Unannotated Transcripts

Some SEPs may have been overlooked (false negatives) in our analysis of RefSeq transcripts due to the presence of RNAs in K562 cells that are not annotated in the RefSeq database. To account for such RNAs we also analyzed the LC-MS/MS peptidomics data using a second custom database derived from K562 RNA-Seq data. Furthermore, recognizing that recent ribosome profiling studies identified a number of

sORFs within the pool of long intergenic non-coding RNAs (lincRNAs) in mouse [2], we generated an extensive catalog of K562 lincRNAs by applying a previously described lincRNA-calling pipeline[28] to our RNA-Seq data and searched the corresponding protein database against our data sets. We applied the same stringent criteria for scoring and assessing peptide-spectral matches, and eliminating peptides with fewer than two differences from annotated proteins; we also eliminated any peptides of fewer than 8 amino acids in order to further reduce false positives. These analyses yielded an additional 54 SEPs.

Combining the RefSeq and RNA-Seq results, we discovered 90 unannotated SEPs, four of which were previously reported and thus served as positive controls[20], and 86 of which are novel (Figure 3.3 C). The average length of each peptide identified using this approach was 13-14 amino acids and 90% of the peptides were longer than 18 amino acids, which supports the use of trypsin to generate an ideal LC-MS/MS sample for SEP discovery. This is the largest number of SEPs ever reported in a single study and increases the total number of known human SEPs[18,20] by ~18-fold, demonstrating the superior coverage afforded by our approach. Interestingly, analysis of the evolutionary conservation of the SEPs across 29 mammalian species suggested that SEPs are more conserved than introns, but not as conserved as known coding genes[29] (Figure. 3.4).

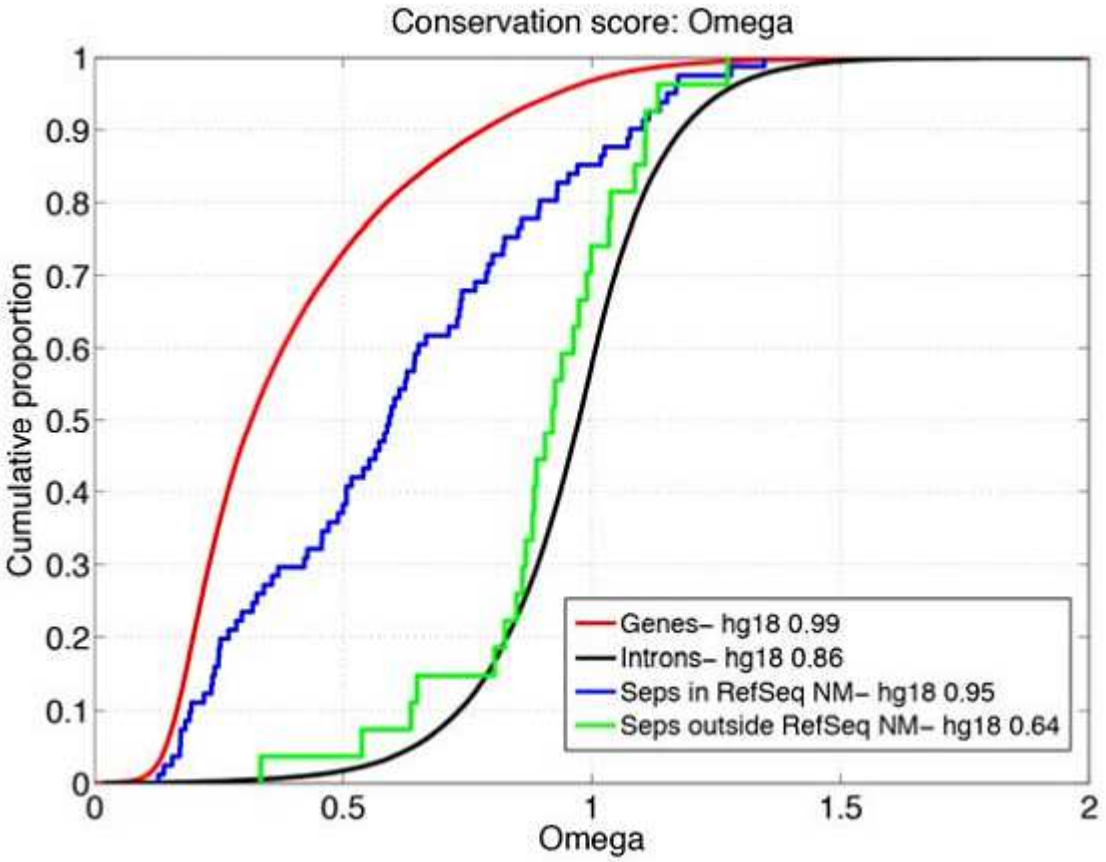


Figure 3.4: Conservation of sORFs. sORFs are conserved more than introns, but less than known coding genes. Higher omega indicates less conservation.

3.4 SEP Translation is Initiated at Non-AUG Codons

Because we performed mass spectrometry on trypsin-digested samples, we do not obtain full protein-level SEP sequence coverage, and in particular do not directly observe the N terminus. We therefore assigned the likely start codon for each SEP in order to determine their lengths. When present, an upstream in-frame AUG was assumed to be the initiation codon. If no upstream AUG was present, the initiation codon was assigned to an in-frame near-cognate non-AUG codon embedded within a Kozak-consensus sequence [30]. In a few cases, neither of these conditions was met, so

the codon immediately following an upstream stop codon was used to determine maximal SEP length.

Using this approach, we determined the SEPs to be 18-149 amino acids long, with the majority (~ 80%) being <100 amino acids (Figure 3.3 C). If we take a more conservative approach by using an AUG-to-stop or upstream-stop-to-stop, we obtain similar SEP length distribution and retain our smallest SEPs. As the shortest human SEP previously identified by mass spectrometry was 88 amino acids long[20], it is clear that our approach provides superior coverage of small SEPs. This is significant because many previously characterized, functional SEPs in other species are under 50 amino acids[10,16-18].

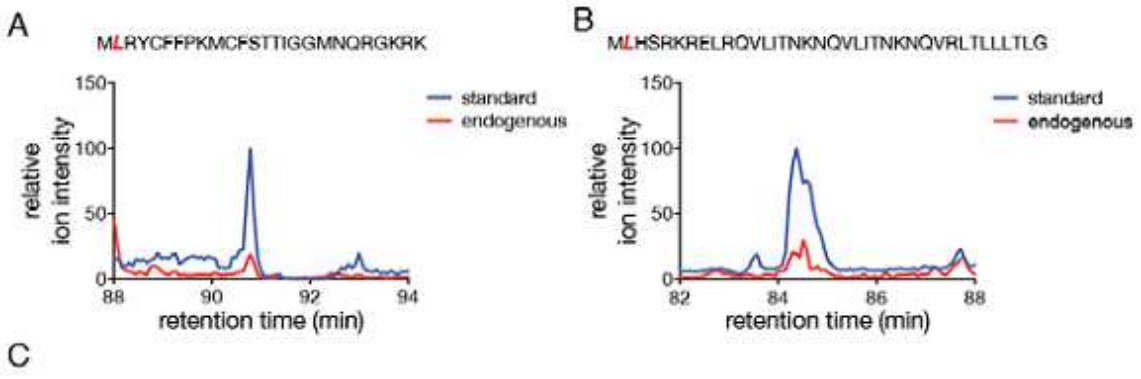
Another interesting feature of our results is the preponderance of non-canonical translation start sites: 57% of the detected SEPs do not initiate at AUG codons (Figure 3.3 D). This finding is consistent with the results of ribosome profiling experiments in mouse, which indicate that, globally, most ORFs contain non-AUG start sites[2]. Below we obtain data demonstrating that these non-AUG sites are the actual initiation codons of the sORFs.

3.5 Supporting SEP length assignments

We used two approaches to gain additional insight into the lengths of our SEPs. First, rather than relying solely on a molecular weight cutoff filter we decided to use polyacrylamide gel electrophoresis (PAGE) to better separate the K562 lysate into different molecular weight fractions. PAGE can be used as a molecular weight fractionation method prior to proteomics and this approach has successfully been used

to study proteolysis[31]. With SEPs, PAGE would provide a tighter molecular weight range, which would support the assigned lengths of the SEPs. Indeed, analysis of the ~10-15 kDa portion of the K562 found SEPs that we had identified as being 90-120 amino acids in length, supporting that these SEPs are intact in these cells which would lead them to migrate at ~10-15 kDa. Importantly, for some of these SEPs we also find additional peptides from the SEP to provide even greater confidence in the SEP assignments.

We still needed to demonstrate that full-length SEPs are present in K562 lysates and therefore we elected to perform an isotope-dilution mass spectrometry (IDMS) experiment with chemically synthesized full-length SEPs. Specifically, we prepared two SEPs, M***L***HSRKRELRQVLITNKNQVLITNKQVRLTLLLTLG and M***L***RCFFPKMCFSTTIGGMNQRGKRK, with a deuterated leucine (d10-Leu, amino acid that is bold, red and in italics). These two peptides were then added to K562 lysate and the sample was analyzed by LC-MS. These peptides co-eluted with peptides from the sample with the correct mass for the natural SEPs (Figure 3.5). Due to the high charge state of the peptides (+5 ions) the tandem MS (CID) was not informative, which led us to use additional methods for confirmation including IDMS of trypsin fragments and cellular imaging experiments. Our current instrumentation configuration is not designed to easily measure full-length SEPs directly from lysates, however, other mass spectrometry methods including top-down proteomics[32] and high-resolution mass spectrometry approaches for peptide detection[33], should enable the discovery and/or validation of full-length SEPs in the future.



C

Sequence	Predicted Ion Mass	Observed Ion Mass
MLHSRKRELRQVLITNKNQVRLTLLLTLG	689.61525^{5+}	689.62^{5+}
MLHSRKRELRQVLITNKNQVRLTLLLTLG	691.62645^{5+}	691.62^{5+}
MLRYC*FFPKMC*FSTTIGGMNQRGKRK	643.72013^{5+}	643.72^{5+}
MLRYC*FFPKMC*FSTTIGGMNQRGKRK	645.73133^{5+}	645.73^{5+}

Figure 3.5: Isotope dilution mass spectrometry (IDMS) of full length deuterated and endogenous SEPs. Coelution demonstrates correct assignment of the full length SEP from the detected peptide. (A) Coelution of MLRYCFFPKMCFSTTIGGMNQRGKRK with a synthetic peptide and coelution of MLHSRKRKELRQVLITNKNQVLITNKNQVRLTLLLTLG (B) demonstrate correct identification of the full length SEP. (C) Predicted and observed precursor ion masses of the synthetic and endogenous peptides. The CID of these peptides was uninterpretable due to the length of these peptides, but coelution indicates correct identification.

3.6 Cellular Concentrations of SEPs

We wished to explore the biological properties of SEPs. First, we examined the cellular concentrations (K562 cells) of three selected SEPs (ASNSD1-SEP, PHF19-SEP and H2AFx-SEP) using isotope dilution mass spectrometry[34]. (We refer to SEPs by appending "-SEP" to the name of the annotated CDS nearest the sORF; the sORF is given the same name but italicized.) These SEPs were found at concentrations between 10 and 2000 copies per cell. Thus, based on previous estimates of protein copy numbers, SEPs are found at concentrations well within the range of typical cellular

proteins[35-37]. We further note that the MS/MS spectra from the synthetic standards used in these experiments were nearly identical to those produced from the endogenous peptide and eluted at the same retention time as same, thus confirming these identifications (Figure 3.6)

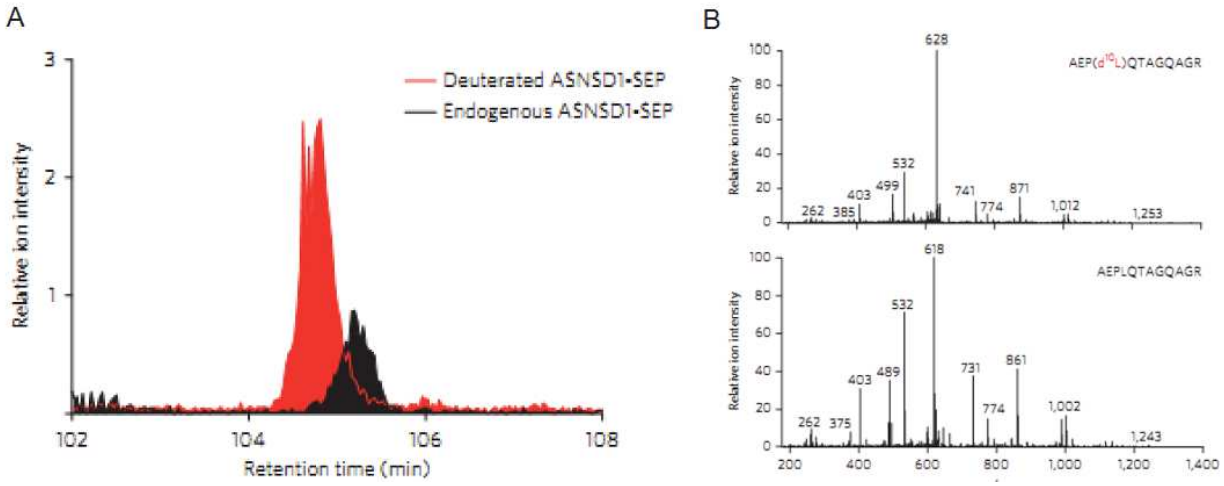


Figure 3.6: SEP quantification. We synthesized deuterated variants of tryptic SEPs. (A) Upon preparation of the K562 peptidome, deuterated SEPs were added exogenously, and the entire mixture was subjected to LC-MS. SEPs were then quantified by comparing the peak areas of the deuterated and endogenous peptide. As the concentration of the deuterated SEP is known, this allowed for the absolute quantification of the SEP. Coelution of the endogenous and deuterated peptides confrims the identification of the endogenous SEP peptide. (B) Matching MS/MS spectra of the heavy and light peptide confirm sequence assignment. 10 Da shifts in some of the MS/MS peaks are due to the presence of the deuterated leucine.

3.7 Heterologous Expression of SEPs

We tested whether the implicated RNA transcripts and sORFs were competent to produce SEPs. Constructs were designed to produce full-length mRNAs, including 5' and 3' UTRs, that matched those in the RefSeq database[38]. We selected sORFs in the 5'-UTR, the 3'-UTR, or frameshifted within the CDS, and encoded a FLAG epitope tag at the 3'-end of each sORF (so that initiation is unperturbed). The uORFs *ASNSD1-SEP*, *PHF19-SEP, DNLZ-SEP, EIF5-SEP, FRAT2-SEP*, *YTHDF3-SEP, CCNA2-SEP,*

DRAP1-SEP, TRIP6-SEP, and *C7ORF47-SEP* all produced cytoplasmically localized polypeptides, as detected by anti-FLAG immunofluorescence in transfected HEK293T cells (Figure 3.7 and Figure 3.8). Most importantly, the fact that *FRAT2-SEP, YTHDF3-SEP*, *CCNA2-SEP, DRAP1-SEP, TRIP6-SEP, C7ORF47-SEP,* which do not have any upstream in frame AUG codons, produced SEPs verifies that sORFs with non-AUG start codons are translated (Figure 3.7 and Figure 3.8).

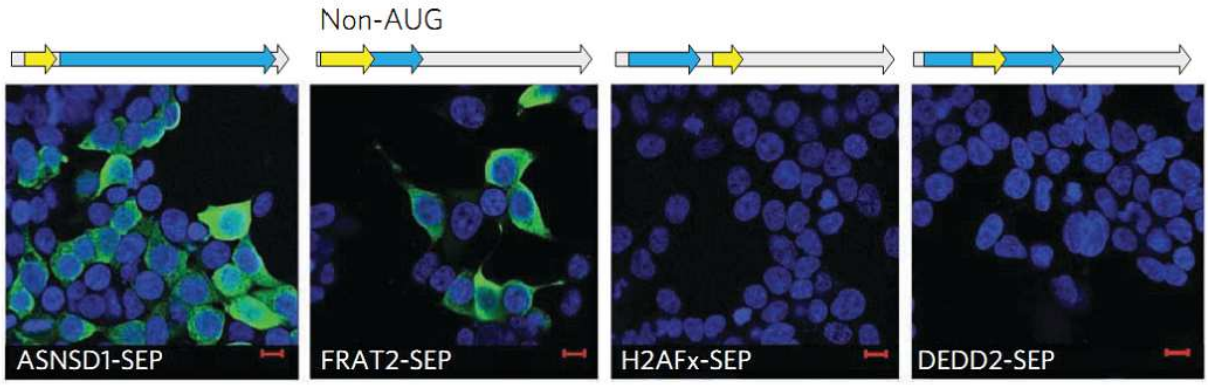


Figure 3.7 Expression of SEPs from their endogenous RNAs. HEK293T cells were transiently transfected with a cDNA construct containing the full length RefSeq mRNA sequence to which SEPs were assigned. The sORF was flagged at the C-terminus. Expression was examined by immunofluorescence (green) and nuclei were stained with Hoescht or Dapi (blue). SEP expression from sORFs originating in the 5'UTR could be observed. sORF originating in the CDS or 3'UTR did not appear to result in SEP expression. However, RNA-Seq data indicated that at least some of these had transcript variants where the sORF was the first AUG on the transcript.

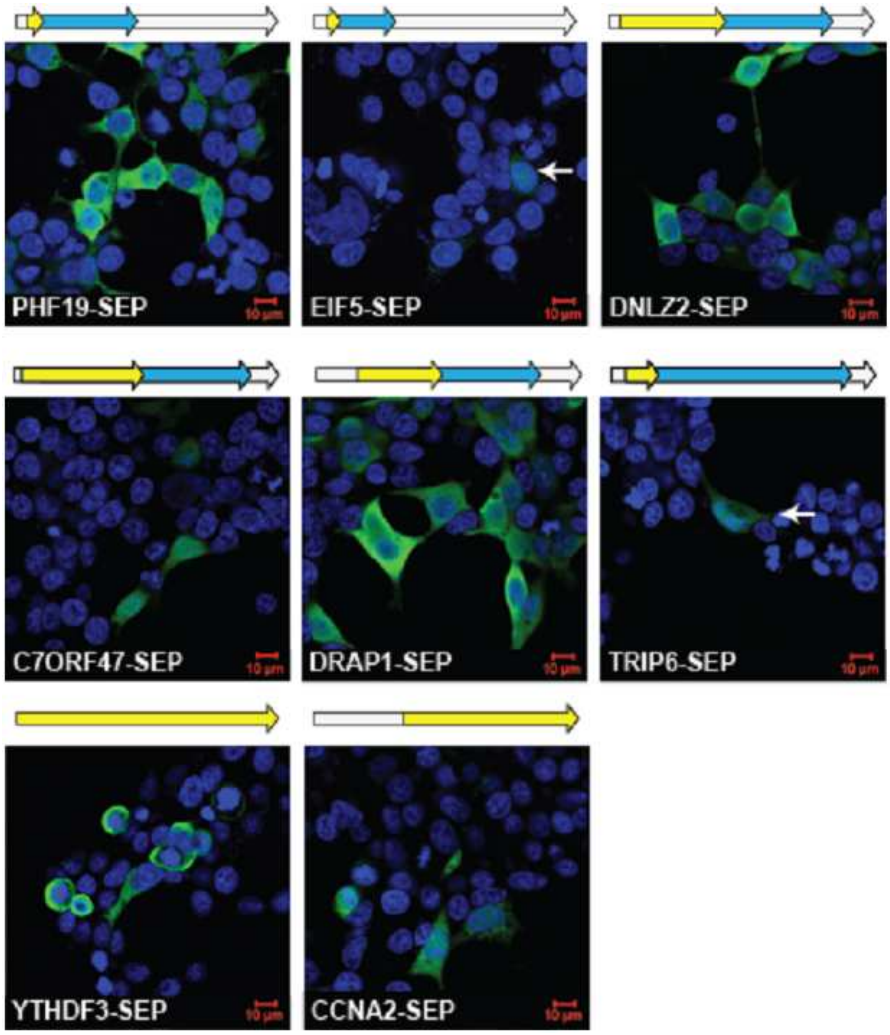


Figure 3.8: Validation of SEP expression. An additional pool of SEPs were expressed from their Refseq transcripts in order to further validate SEP expression.

By contrast, the *DEDD2-SEP* sORF was not translated from the full-length RefSeq construct. *DEDD2-SEP* is frameshifted deep within the main CDS of the DEDD2 transcript, so according to the scanning model of translation[39] it is not expected that this downstream sORF would be translated (Figure 3.7). One possible explanation for our observation of the DEDD2-SEP is that it is translated from a splice variant of the DEDD2 RNA that is present in K562 cells, but is not in RefSeq. In support of this

hypothesis, we identified a truncated DEDD2 mRNA in the RNA-seq data wherein the first start codon is that of the *DEDD2-SEP* sORF (Figure 3.9). The 3'-UTR-embedded H2AFx-SEP was similarly not translated from the full-length mRNA construct; however, we were not able to clearly identify a truncated version of the H2AFx transcript in the K562 RNA-seq data. It is possible that a truncated H2AFx mRNA variant is present in K562 cells but is not detectable or not resolvable from the full-length H2AFx transcript.

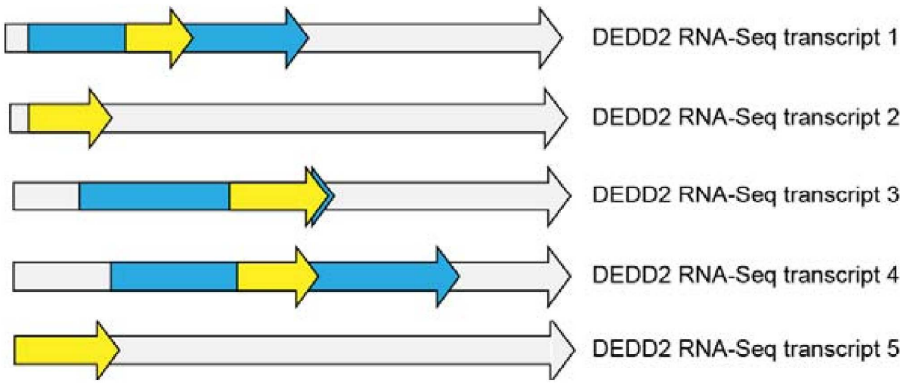


Figure 3.9: DEDD2 has truncated transcript variants. RNA-Seq data indicates there are multiple transcript variants of DEDD2. In some of these variants the 5' end of the transcript is truncated, and the transcript lacks the DEDD2 CDS start codon. In these transcripts the sORF initiation codon is the first AUG on the transcript.

3.8 SEPs Exhibit Subcellular Localization

We subcloned expression constructs for FLAG-tagged DEDD2-SEP and H2AFx-SEP to determine whether these SEPs are stable. The *H2AFx-SEP* sORF produced a cytoplasmic polypeptide in HEK293T cells (Figure 3.10).

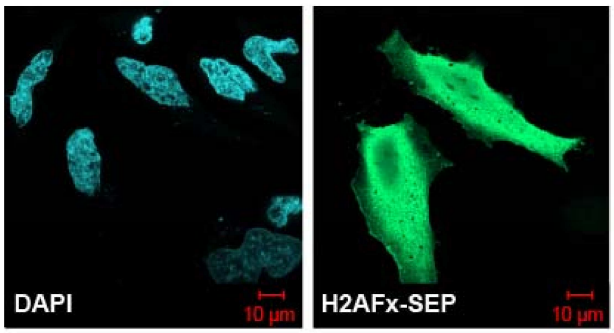


Figure 3.10: H2AFx SEP is cytoplasmic. Expression of the H2AFx flag tagged sORF demonstrated H2AFx-SEP was a cytoplasmic polypeptide in HEK293T cells.

Interestingly, DEDD2-SEP localizes to mitochondria in HEK293T, mouse embryonic fibroblast (MEF), and COS7 cells, as demonstrated by co-localization with the mitochondrial marker MitoTracker Red (Figure 3.11). The N-terminus of DEDD2-SEP is predicted to contain a mitochondrial import signal[40]. Sequence-dependent trafficking and subcellular localization of SEPs could therefore be general phenomena related to their biological activities.

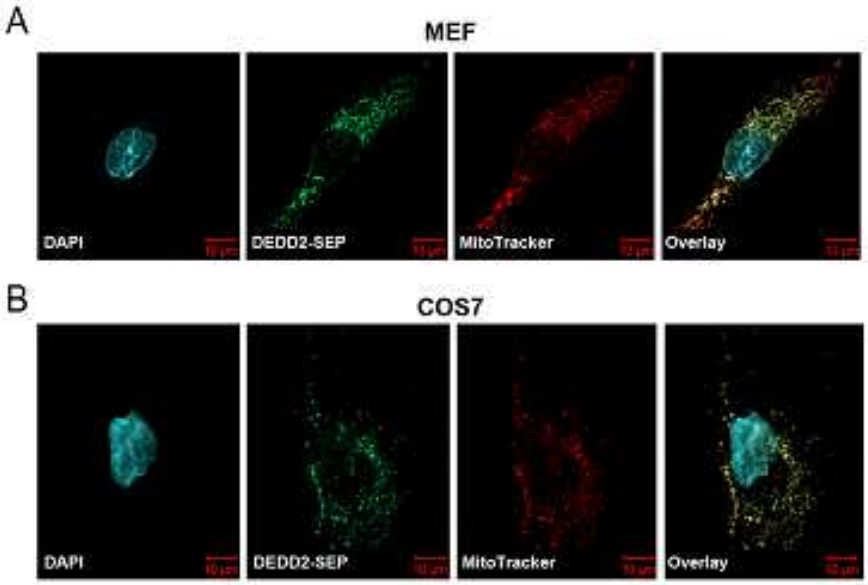


Figure 3.11: DEDD2-SEP localizes to the mitochondria. DEDD2-SEP-sORF was expressed in HEK293T, MEF, and COS7 cells in order to examine its localization (green). Costaining with MitoTracker (red) indicated DEDD2-SEP localized to the mitochondria.

3.9 Non-AUG Start Codons Enable Bicistronic Expression

Since such a large proportion of SEPs putatively initiate at non-AUG sites, we wanted to rigorously identify the alternate start codon of one these sORFs. C-terminally FLAG-tagged FRAT2-SEP was expressed from the full-length mRNA construct in HEK293T cells and immunoprecipitated; mass spectrometry of the purified protein (Figure 3.11) was consistent with initiation at an ACG triplet embedded within a Kozak consensus sequence[30] (Figure 3.12).

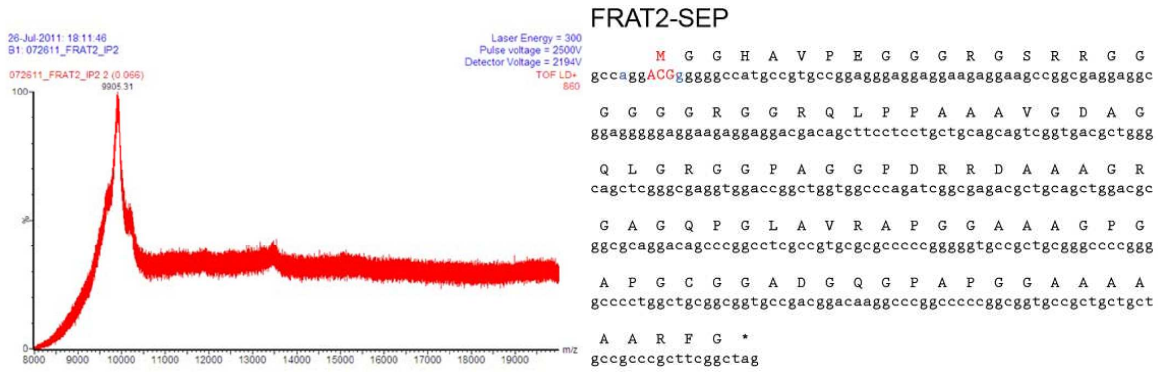


Figure 3.12: FRAT2 ACG misprimes for methionine. Immunoprecipitation of FRAT2 SEP followed by MALDI indicated FRAT2 has initiates with a methionine despite an ACG initiation codon.

Mutating the ACG to an ATG resulted in increased FRAT2-SEP translation while deletion of this ACG abolished FRAT2-SEP production, as assessed by Western blotting, thus confirming our assignment (Figure 3.13). In addition, mutation of the Kozak consensus sequence to less favorable residues led to markedly lower FRAT2-SEP expression, which demonstrates the importance of the Kozak sequence at non-AUG initiation sites.

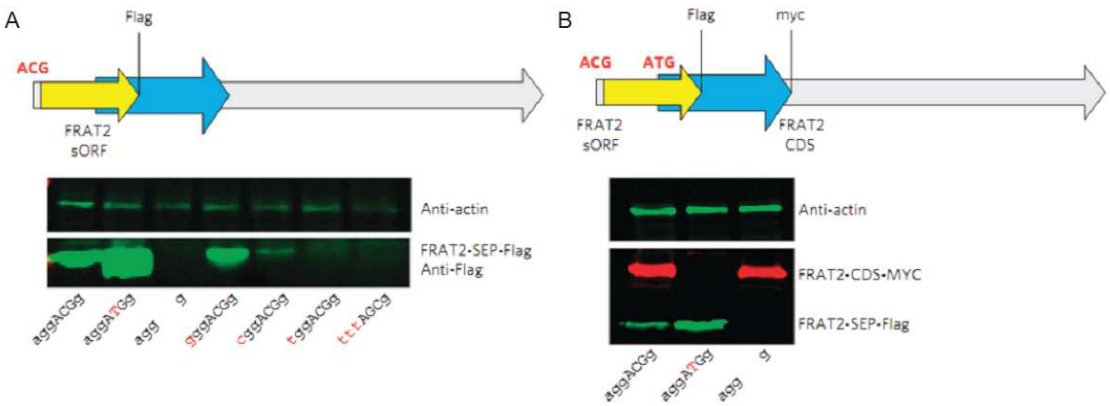


Figure 3.13 FRAT2 mRNA is bicistronic. (A) The FRAT2 cDNA expression construct, with a Flag epitope tag appended at the C terminus of the FRAT2-SEP sORF was subjected to site-directed mutagenesis to probe the identity of the sORF start codon by expression in HEK293T cells followed by western blotting. Below the immunoblot, the sORF Kozak and start codon sequences of the expressed construct are shown, with the start codon shown in uppercase and sites of mutation highlighted in red. Conversion of the putative ACG start codon to an ATG resulted in higher expression (lane 3). In addition, perturbation of the kozak sequence (lane 4-7) revealed the importance of context when using non-AUG codons, as substitution of less favorable residues resulted in lower FRAT2-SEP expression. Equal loading was demonstrated with actin specific immunoblotting. (B) Epitope tagging of the sORF and CDS of the FRAT2 mRNA demonstrates that the FRAT2 mRNA is bicistronic. The FRAT2 CDS was c-myc tagged, and the FRAT2-SEP was Flag tagged. Conversion of the FRAT2-SEP initiation codon from ACG to ATG ablates expression of the downstream FRAT2 CDS, indicating the importance of alternate start codons for polycistronic expression. RNA maps are not to scale.

The scanning model of translation provided an explanation as to why the DEDD2 mRNA is not bi-cistronic; we hypothesized that upstream alternate start codons could provide a mechanism to promote polycistronic gene expression via leaky scanning. To test whether FRAT2 mRNA is bi-cistronic, we prepared a FRAT2 construct where the SEP and the downstream CDS were tagged with different epitopes (Figure 3.13), permitting their simultaneous detection by immunoblotting with two antibodies. We found that the FRAT2 RNA is bi-cistronic, as FRAT2 and FRAT2-SEP are both expressed (Figure 3.13). Remarkably, mutation of the ACG start codon of the *FRAT2*-

SEP to an ATG increases FRAT2-SEP expression, but also completely eliminates the expression of FRAT2 protein, revealing that the translation of the downstream cistron absolutely requires leaky upstream initiation. Therefore, this experiment indicated that an upstream non-AUG initiation codon is necessary for efficient polycistronic gene expression.

While we attribute FRAT2-SEP translation and bi-cistronic expression to alternate start codon use, we note that another interesting mechanistic possibility for FRAT2-SEP translation is partial (or incomplete) RNA editing, which could modify the ACG to AUG post-transcriptionally. The role of RNA editing in generating sORF start codons at the RNA level could be studied in the future via genetic knockout of the enzymes responsible for this activity[41].

3.10 A Small Subset of lincRNAs encode SEPs

Another intriguing feature of these experiments was the discovery of SEPs encoded by lincRNAs. lincRNAs have emerged as an important class of regulatory molecules with intrinsic biological functions (e.g., hotair, xist)[42,43]. Ribosome profiling experiments in mouse cells indicate the presence of translated sORFs on nearly half of the lincRNAs analyzed[2], which is much higher than expected[42,44,45]. By contrast, our peptidomics analysis identified 8 SEP-encoding lincRNAs, which represents just 0.4% of the1866 lincRNAs detected in our RNA-seq analysis of K562.

This disparity may result from a number of factors, including false positive identifications by ribosome profiling techniques [3,4]. Indeed, many of these IDs were later found to be erroneous by more thorough re-analysis of ribosome profiling data[4]. Additionally, ribosome profiling may identify rare translational events that do not

generate enough protein to be detected by LC-MS/MS, since mass spectrometry is biased towards the detection of more abundant peptides[46]. It is also possible that some of the sORFs identified by ribosome profiling may produce polypeptides that are rapidly degraded and therefore would be undetectable using any analytical approach. Future work coupling ribosome profiling with mass spectrometry should help resolve these questions and provide a better understanding of the factors governing SEP expression.

3.11 Conclusion

In contrast to previous attempts to use mass spectrometry to discover unannotated human coding sequences, we successfully access the pool of SEPs that are under 50 amino acids in length. This is unprecedented for a global discovery technique and is a crucial step towards understanding the biology of these molecules, for many of the known SEPs[16-18] are below this size threshold. Moreover, the unbiased discovery of SEPs also provided insights into protein translation through the characterization of non-AUG codons and validation of mammalian polycistronic gene expression. Taken together, these findings provide the strongest evidence to date that coding sORFs constitute a significant human gene class. Moreover, due to the bias of mass spectrometry for more abundant species[46], which limits the scope of our technique to the most highly expressed SEPs, and our conservative identification criteria it is probable that there are many more as-yet-undiscovered human SEPs. Thus, we believe we have only begun to explore the breadth and diversity of this new family of polypeptides.

3.12 Methods

Cloning and mutagenesis:

DNA constructs were prepared by standard ligation, Quikchange, or inverse PCR techniques. Human cDNA clones were obtained from Open Biosystems and subcloned into pcDNA3, which uses a CMV promoter. Gene synthesis was performed by DNA2.0. Plasmid sequences are publicly available upon request. We note that the YTHDF3-SEP construct consisted of the 5'-UTR putatively encoding the SEP only, obtained via gene synthesis because a full-length cDNA construct with an intact 5'-UTR was not commercially available.

Cell culture:

Cells were grown at 37ºC under an atmosphere of 5% CO_2. HEK293T, HeLa, COS7 and MEF cells were grown in high-glucose DMEM supplemented with L-glutamine, 10% fetal bovine serum, penicillin and streptomycin. K562 cells were maintained at a density of 1-10 x 10^5 cells/mL in RPMI1640 media with 10% fetal bovine serum, penicillin and streptomycin.

Isolation and processing of polypeptides:

Aliquots of 3 x 10^7 growing K562 cells were placed in 1.5 ml Protein LoBind Tubes (Eppendorf), washed three times with PBS, pelleted and stored at -80 °C. Boiling water (500 µl) was added directly to the frozen cell pellets and the samples were then boiled for 20 minutes to eliminate proteolytic activity [21,23]. After cooling to room temperature, samples were sonicated on ice for 20 bursts at output level 4 with a 40% duty cycle (Branson Sonifier 250; Ultrasonic Convertor). The cell lysate was then brought to 0.25% acetic acid by volume and centrifuged at 20,000 x g for 20 minutes at 4°C. The

supernatant was sent through a 30 kD or 10 kD molecular weight cut-off (MWCO) filter (Modified PES Centrifugal Filter, VWR). The mix of small proteins and peptides in the flow-through was evaluated for protein content by BSA assay and then evaporated to dryness at low temperature in a SpeedVac. Pellets were re-suspended in 50 µl of 25mM TCEP in 50mM NH_4HCO_3 (pH=8) and incubated at 37 °C for 1 hour. The reaction was cooled to room temperature before 50 µl of a 50 mM iodoacetamide solution in 50 mM NH_4HCO_3. This solution was incubated in the dark for 1 hour. Samples were then dissolved in a 50 mM NH_4HCO_3 solution of 20 µg/µl trypsin (Promega) to a final protein to enzyme mass ratio of 50:1. This reaction was incubated at 37 °C for 16 hours, cooled to room temperature and then quenched by adding neat formic acid to 5% by volume. The digested peptide mix was then bound to a C18 Sep Pak cartridge (HLB, 1cm^3; 30mg, Oasis), washed thoroughly with water and eluted with 1:1 acetonitrile/water. The eluate was evaporated to dryness at low temperature on a SpeedVac.

Offline electrostatic repulsion-hydrophilic interaction chromatography (ERLIC) fractionation of polypeptide fraction.

To simplify the sample and thereby improve detection sensitivity in the subsequent LC-MS/MS analysis, we separated the processed peptide mix by ERLIC[47,48]. ERLIC was performed using a PolyWax LP column (200 x 2.1 mm, 5µm, 300Å; PolyLC Inc.) connected to an Agilent Technologies 1200 Series HPLC equipped with a degasser and automatic fraction collector. All runs were performed at a flow rate of 0.3 ml/min and ultraviolet absorption was measured at a wavelength of 210 nm. Forty (30 kD sample) or 25 (10 kD sample) fractions were collected over a 70 minute gradient beginning with 0.1% acetic acid in 90% acetonitrile (aq.) and ending with 0.1% formic acid in 30%

acetonitrile (aq.). The fractions were then evaporated to dryness on a SpeedVac and dissolved in 15 µl 0.1% formic acid (aq.) in preparation for LC-MS/MS analysis.

LC-MS/MS analysis. Samples were injected onto a NanoAcquity HPLC system (Waters) equipped with a 5 cm x 100 µm capillary trapping column (New Objective) packed with 5 um C18 AQUA beads (Waters) and a PicoFrit SELF/P analytical column (15 µm tip, 25 cm length, New Objective) packed with 3 µm C18 AQUA beads (Waters) and separated over a 90 minute gradient at 200 nl/min. This HPLC system was online with an LTQ Orbitrap Velos (Thermo Scientific) instrument, which collected full MS (dynamic exclusion) and tandem MS (Top 20) data over 375-1600 m/z with 60,000 resolving power.

Data processing:

The acquired MS/MS spectra were analyzed with the SEQUEST algorithm using a database derived from 6-frame (forward and reverse) translation of RefSeq (National Center for Biotechnology Information) mRNA transcripts or 3-frame (forward only) translation of a transcriptome assembly generated by Cufflinks[49] using RNA-Seq data from the K562 cell line (data acquisition described below). The search was performed with the following parameters: variable modifications, oxidation (Met), N-acetylation; semitryptic requirement; maximum missed cleavages: 2; precursor mass tolerance: 20 ppm; and fragment mass tolerance: 0.7 Da. Search results were filtered such that the estimated false discovery rate of the remaining results was 1%. The Sf score is the final score for protein identification by the Proteomics Browser software based on a combination of the preliminary score, the cross-correlation and the differential between the scores for the highest scoring protein and second highest scoring protein[27].

Identified peptides were searched against the Uniprot human protein database using a string-searching algorithm. Peptides found to be identical to fragments of annotated proteins were eliminated from the SEP candidate pool. The remaining peptides were searched against non-redundant protein sequences using the Basic Local Alignment Search Tool (BLAST). Any peptides found to be less than two amino acids different from the nearest protein match (i.e., identical or different by one amino acid) were discarded.

The spectra of the remaining peptides were subjected to a rigorous manual validation procedure: spectra were rejected if they had a precursor mass error of >5 ppm, if they lacked a sequence tag of 5 consecutive b- or y-ions, if they had more than one missed cleavage, or if they lacked sufficient sequence coverage to differentiate from the nearest annotated protein. Finally, peptides under 8 amino acids in length were discarded in order to further minimize false positive identifications.

RNA-Seq library preparation, alignment, and transcriptome assembly:

Two types of cDNA libraries were generated from K-562 RNA (Ambion). In the first experiment, we used 50 nanograms of polyA$^+$ RNA to create standard, non-strand-specific cDNA libraries with paired-end adaptors as previously described [50] and sequenced it on one lane of an Illumina Genome Analyzer IIa machine. In the second experiment, we used eight different amounts of total RNA (30 ng, 100 ng, 250 ng, 500 ng, 1000ng, 3000 ng, and 10,000 ng) to create cDNA libraries with paired-end, indexed adaptors following the instructions for the Illumina TruSeq RNA sample prep kit, except that we used SuperScript III instead of SuperScript II and optimized PCR cycle number. These libraries were sequenced on a single lane of a HiSeq2000 machine. RNA-Seq

reads were aligned to the human genome (Hg19 assembly) using TopHat [version V1.1.4;[51]] and transcriptome assembly was performed using Cufflinks [version V1.0.0;[49]]. lincRNAs were called based on a lincRNA-calling pipeline as previously described[28]. The transcriptome data is deposited on GEO (GSE34740).

Peptide synthesis, purification and concentration determination**:**

Automated (PS3 Protein Technology, Inc.) solid-phase peptide synthesis was carried out using Fmoc amino acids. Crude peptides were HPLC (Shimadzu)-purified using a C18 column (150 mm × 20 mm, 10 μm particle size, Higgins Analytical). The mobile phase was adjusted for each peptide; buffer A was 99% H_2O, 1% acetonitrile, and 0.1% TFA; buffer B was 90% acetonitrile, 10% H_2O, and 0.07% TFA). Pure fractions were identified by MALDI-MS analysis, combined, and lyophilized. Peptide concentrations were determined by amino acid analysis (AlBio Tech).

Absolute quantification of SEPs:

Isotope dilution mass spectrometry (IDMS) [34] was used to determine the concentration of SEPs in K562 cells. All samples for this experiment were prepared by adding known amounts of heavy isotope-labeled peptides corresponding the detected fragment of the SEP of interest to a K562 cell pellet (10^7 cells) just before isolation of the polypeptides from these cells. The preparation of these samples was identical to that described above except that no ERLIC separation was done. The first step of the quantification procedure was to prepare a set of samples where each sample contained a different but known amount (1 fmol, 10 fmol, 50 fmol, 100 fmol, 500 fmol, 1 pmol or 10 pmol) of the heavy-labeled counterpart peptide. These samples were then analyzed by a selected ion monitoring (SIM) method on the previously described LC-MS/MS system and the

resulting data was analyzed using Xcaliber 2.0 (Thermo Scientific). The areas of the peaks corresponding to the endogenous and isotope-labeled peptides were compared to determine the approximate concentration of the SEP and a standard curve was generated to verify that the quantity of the SEP fragment was within the linear range of the mass spectrometer. A second set of samples that each contained an amount of isotope-labeled peptide that was within the linear range of the instrument and within an order of magnitude of the amount of the corresponding endogenous peptide in the cells was then prepared (N=4) and analyzed as described. The results of this experiment were used to determine the absolute cellular concentration of the selected SEPs.

Imaging SEPs by immunofluorescence:

HeLa, COS7, and MEF cells were grown to 80% confluency on glass coverslips in 48-well plates; HEK293T cells were grown to 50-75% confluency on fibronectin (Millipore)-coated glass coverslips in 48-well plates. Cells were transfected in Opti-MEM (Invitrogen) with 250 ng plasmid DNA using Lipofectamine 2000 (Invitrogen) according to the manufacturer's instructions. 24 hours after transfection, cells were fixed with 4% formalin in phosphate buffered saline (PBS) for 10 minutes at room temperature, and then permeabilized with methanol at -20ºC for 10 minutes. Fixed cells were blocked with blocking buffer (3% BSA in PBS with 0.5% Tween-20), then incubated overnight at 4ºC with anti-FLAG M2 antibody (Sigma) diluted 1:1000 in blocking buffer. After washing 3x with PBS, cells were then stained for one hour at room temperature with goat anti-mouse AlexaFluor 488 conjugate (Invitrogen) diluted 1:1000 in blocking buffer. Cells were washed 3x with PBS, post-fixed with 4% formalin for 10 minutes at room temperature, then counterstained with a final concentration of 270 ng/mL Hoescht

33258 (Invitrogen) in PBS for 15 minutes at room temperature. Cells were then imaged in PBS in glass-bottom imaging dishes (Matek Corp.). For mitochondrial co-localization analysis, transfected cells were treated with MitoTracker Red CMXRos (Invitrogen) at a final concentration of 100 nM in PBS at 37ºC for 15 minutes, washed once with PBS, then fixed with formalin and methanol and immunostained as described above.

Images were acquired in the Harvard Center for Biological Imaging on a Zeiss LSM 510 inverted confocal microscope with the following lasers: 405 Diode, 488 (458,477,514) Argon, 543 HeNe and 633 HeNe. Image acquisition was with either AIM or Zen software. Images were acquired with a 60x oil immersion objective.

Determination of the FRAT2-SEP start codon by immunoprecipitation and MALDI-MS:

COS7 and HEK293T cells were grown in 10-cm dishes to 75% confluency, then transfected with 10 µg plasmid DNA using Lipofectamine 2000 according to the manufacturer's instructions. 24 hours after transfection, cells were harvested by scraping and washed 3x with PBS. Cells were lysed in 400 µL Triton lysis buffer (1% Triton X-100 in Tris-buffered saline (TBS) with Roche Complete Mini Protease Inhibitor added) on ice for 15 minutes, then lysates were clarified by centrifugation at 16,100 x g for 20 minutes at 4ºC. Clarified lysates were added to 50 µL of PBS-washed anti-FLAG M2 agarose resin (Sigma) and rotated at 4ºC for 1 hour. Beads were washed 6x with TBS-T (Tris-buffered saline with 0.05% Tween-20). To elute bound proteins, 50 uL of 100 ug/mL 3x FLAG peptide (Sigma) in TBS-T was added to the resin and rotated at 4ºC for 20 minutes. Eluates were stored at -80ºC until further analysis.

For MALDI-MS analysis, the entire protein sample was desalted using a C18 Sep Pak cartridge (HLB, $1cm^3$; 30mg, Oasis) and eluted in 50% acetonitrile. The sample was

dried in a SpeedVac, and then dissolved in a final volume of 10 µL mass spectrometry-grade water (Burdick & Jackson). This solution (1 µL) was mixed with matrix (α-cyano-4-hydroxycinnamic acid in 50% acetonitrile, 1 µL) on a stainless steel MALDI plate and air-dried. Data were acquired on a Waters MALDI micro MX instrument operated in linear positive mode. Instrument control and spectral acquisition were with MassLynx software.

Confirmation of the FRAT2-SEP initiation codon, Kozak sequence, and bicistronic expression by immunoblotting:

HEK293T cells were grown to 75% confluency in 6-well plates, then transfected with 10 µg plasmid DNA using Lipofectamine 2000 according to the manufacturer's instructions. Cells were harvested by vigorous pipetting and lysed in 100 µL Triton lysis buffer. Samples of clarified lysate (20 µL) were mixed with SDS-PAGE loading buffer, boiled, and electrophoresed on 4-20% Tris-HCl gels (Bio-Rad). Two replicate gels were run. Proteins were transferred to nitrocellulose (0.20 µm pore size, Thermo Scientific) and immunoblots were probed with anti-FLAG M2 antibody (Sigma) followed by goat anti-mouse IR dye 800 conjugate (LICOR). For bicistronic expression assays, immunoblots were probed with a mixture of rabbit anti-c-myc antibody (Sigma) and anti-FLAG M2, followed by a mixture of goat anti-mouse IR dye 800 and goat anti-rabbit IR dye 680 (LICOR). A replica immunoblot was probed with mouse anti-β-actin followed by goat anti-mouse IR dye 800. Antibodies were diluted 1:2000 in Rockland Immunochemicals fluorescent blocking buffer. Infrared imaging was performed on a LICOR Odyssey instrument.

3.13 References

(1) Frith, M. C.; Forrest, A. R.; Nourbakhsh, E.; Pang, K. C.; Kai, C.; Kawai, J.; Carninci, P.; Hayashizaki, Y.; Bailey, T. L.; Grimmond, S. M. *PLoS Genet* **2006**, *2*, e52.

(2) Ingolia, N. T.; Lareau, L. F.; Weissman, J. S. *Cell* **2011**, *147*, 789.

(3) Zhang, F.; Hinnebusch, A. G. *Nucleic Acids Res* **2011**, *39*, 3128.

(4) Guttman, M.; Russell, P.; Ingolia, N. T.; Weissman, J. S.; Lander, E. S. *Cell* **2013**, *154*, 1.

(5) Calvo, S. E.; Pagliarini, D. J.; Mootha, V. K. *Proc Natl Acad Sci U S A* **2009**, *106*, 7507.

(6) Abastado, J. P.; Miller, P. F.; Hinnebusch, A. G. *New Biol* **1991**, *3*, 511.

(7) Kozak, M. *Cell* **1986**, *47*, 481.

(8) Parola, A. L.; Kobilka, B. K. *J Biol Chem* **1994**, *269*, 4497.

(9) Werner, M.; Feller, A.; Messenguy, F. *Cell* **1987**.

(10) Wadler, C. S.; Vanderpool, C. K. *Proc Natl Acad Sci U S A* **2007**, *104*, 20454.

(11) Jay, G.; Nomura, S.; Anderson, C. W.; Khoury, G. **1981**.

(12) Casson, S. A.; Chilley, P. M.; Topping, J. F.; Evans, I. M.; Souter, M. A.; Lindsey, K. *Plant Cell* **2002**, *14*, 1705.

(13) Rohrig, H.; Schmidt, J.; Miklashevichs, E.; Schell, J.; John, M. *Proc Natl Acad Sci U S A* **2002**, *99*, 1915.

(14) Kastenmayer, J. P.; Ni, L.; Chu, A.; Kitchen, L. E.; Au, W. C.; Yang, H.; Carter, C. D.; Wheeler, D.; Davis, R. W.; Boeke, J. D.; Snyder, M. A.; Basrai, M. A. *Genome Res* **2006**, *16*, 365.

(15) Gleason, C. A.; Liu, Q. L.; Williamson, V. M. *Mol Plant Microbe Interact* **2008**, *21*, 576.

(16) Galindo, M. I.; Pueyo, J. I.; Fouix, S.; Bishop, S. A.; Couso, J. P. *PLoS Biol* **2007**, *5*, e106.

(17) Kondo, T.; Hashimoto, Y.; Kato, K.; Inagaki, S.; Hayashi, S.; Kageyama, Y. *Nat Cell Biol* **2007**, *9*, 660.

(18) Hashimoto, Y.; Niikura, T.; Tajima, H.; Yasukawa, T.; Sudo, H.; Ito, Y.; Kita, Y.; Kawasumi, M.; Kouyama, K.; Doyu, M.; Sobue, G.; Koide, T.; Tsuji, S.; Lang, J.; Kurokawa, K.; Nishimoto, I. *Proc Natl Acad Sci U S A* **2001**, *98*, 6336.

(19) Hemm, M. R.; Paul, B. J.; Schneider, T. D.; Storz, G.; Rudd, K. E. *Molecular Microbiology* **2008**, *70*, 1487.

(20) Oyama, M.; Kozuka-Hata, H.; Suzuki, Y.; Semba, K.; Yamamoto, T.; Sugano, S. *Mol Cell Proteomics* **2007**, *6*, 1000.

(21) Tinoco, A. D.; Tagore, D. M.; Saghatelian, A. *Journal of the American Chemical Society* **2010**, *132*, 3819.

(22) Svensson, M.; Skold, K.; Svenningsson, P.; Andren, P. E. *Journal of proteome research* **2003**, *2*, 213.

(23) Tagore, D. M.; Nolte, W. M.; Neveu, J. M.; Rangel, R.; Guzman-Rojas, L.; Pasqualini, R.; Arap, W.; Lane, W. S.; Saghatelian, A. *Nat Chem Biol* **2009**, *5*, 23.

(24) Swaney, D. L.; Wenger, C. D.; Coon, J. J. *Journal of proteome research* **2010**, *9*, 1323.

(25) Eng, J. K.; McCormack, A. L.; Yates Iii, J. R. *Journal of the American Society for Mass Spectrometry* **1994**, *5*, 976.

(26) Yates, J. R., 3rd; Eng, J. K.; McCormack, A. L.; Schieltz, D. *Analytical chemistry* **1995**, *67*, 1426.

(27) Christofk, H. R.; Vander Heiden, M. G.; Wu, N.; Asara, J. M.; Cantley, L. C. *Nature* **2008**, *452*, 181.

(28) Cabili, M. N.; Trapnell, C.; Goff, L.; Koziol, M.; Tazon-Vega, B.; Regev, A.; Rinn, J. L. *Genes Dev* **2011**, *25*, 1915.

(29) Garber, M.; Guttman, M.; Clamp, M.; Zody, M. C.; Friedman, N.; Xie, X. *Bioinformatics* **2009**, *25*, i54.

(30) Kozak, M. *Cell* **1986**, *44*, 283.

(31) Dix, M. M.; Simon, G. M.; Cravatt, B. F. *Cell* **2008**, *134*, 679.

(32) Tran, J. C.; Zamdborg, L.; Ahlf, D. R.; Lee, J. E.; Catherman, A. D.; Durbin, K. R.; Tipton, J. D.; Vellaichamy, A.; Kellie, J. F.; Li, M.; Wu, C.; Sweet, S. M.; Early, B. P.; Siuti, N.; LeDuc, R. D.; Compton, P. D.; Thomas, P. M.; Kelleher, N. L. *Nature* **2011**, *480*, 254.

(33) Kersten, R. D.; Yang, Y. L.; Xu, Y.; Cimermancic, P.; Nam, S. J.; Fenical, W.; Fischbach, M. A.; Moore, B. S.; Dorrestein, P. C. *Nat Chem Biol* **2011**, *7*, 794.

(34) Keshishian, H.; Addona, T.; Burgess, M.; Kuhn, E.; Carr, S. A. *Mol Cell Proteomics* **2007**, *6*, 2212.

(35) de Godoy, L. M.; Olsen, J. V.; Cox, J.; Nielsen, M. L.; Hubner, N. C.; Frohlich, F.; Walther, T. C.; Mann, M. *Nature* **2008**, *455*, 1251.

(36) Schwanhausser, B.; Busse, D.; Li, N.; Dittmar, G.; Schuchhardt, J.; Wolf, J.; Chen, W.; Selbach, M. *Nature* **2011**, *473*, 337.

(37) Beck, M.; Schmidt, A.; Malmstroem, J.; Claassen, M.; Ori, A.; Szymborska, A.; Herzog, F.; Rinner, O.; Ellenberg, J.; Aebersold, R. *Molecular systems biology* **2011**, *7*, 549.

(38) Pruitt, K. D.; Tatusova, T.; Maglott, D. R. *Nucleic Acids Res* **2007**, *35*, D61.

(39) Hinnebusch, A. G. *Microbiol Mol Biol Rev* **2011**, *75*, 434.

(40) Bendtsen, J. D.; Nielsen, H.; von Heijne, G.; Brunak, S. *J Mol Biol* **2004**, *340*, 783.

(41) Wedekind, J. E.; Dance, G. S.; Sowden, M. P.; Smith, H. C. *Trends in genetics : TIG* **2003**, *19*, 207.

(42) Guttman, M.; Amit, I.; Garber, M.; French, C.; Lin, M. F.; Feldser, D.; Huarte, M.; Zuk, O.; Carey, B. W.; Cassady, J. P.; Cabili, M. N.; Jaenisch, R.; Mikkelsen, T. S.; Jacks, T.; Hacohen, N.; Bernstein, B. E.; Kellis, M.; Regev, A.; Rinn, J. L.; Lander, E. S. *Nature* **2009**, *458*, 223.

(43) Mercer, T. R.; Dinger, M. E.; Mattick, J. S. *Nat Rev Genet* **2009**, *10*, 155.

(44) Guttman, M.; Garber, M.; Levin, J. Z.; Donaghey, J.; Robinson, J.; Adiconis, X.; Fan, L.; Koziol, M. J.; Gnirke, A.; Nusbaum, C.; Rinn, J. L.; Lander, E. S.; Regev, A. *Nature biotechnology* **2010**, *28*, 503.

(45) Khalil, A. M.; Guttman, M.; Huarte, M.; Garber, M.; Raj, A.; Rivea Morales, D.; Thomas, K.; Presser, A.; Bernstein, B. E.; van Oudenaarden, A.; Regev, A.; Lander, E. S.; Rinn, J. L. *Proc Natl Acad Sci U S A* **2009**, *106*, 11667.

(46) Fonslow, B. R.; Carvalho, P. C.; Academia, K.; Freeby, S.; Xu, T.; Nakorchevsky, A.; Paulus, A.; Yates, J. R., 3rd *Journal of proteome research* **2011**, *10*, 3690.

(47) Alpert, A. J. *Analytical chemistry* **2008**, *80*, 62.

(48) Hao, P.; Guo, T.; Li, X.; Adav, S. S.; Yang, J.; Wei, M.; Sze, S. K. *Journal of proteome research* **2010**, *9*, 3520.

(49) Trapnell, C.; Williams, B. A.; Pertea, G.; Mortazavi, A.; Kwan, G.; van Baren, M. J.; Salzberg, S. L.; Wold, B. J.; Pachter, L. *Nature biotechnology* **2010**, *28*, 511.

(50) Levin, J. Z.; Yassour, M.; Adiconis, X.; Nusbaum, C.; Thompson, D. A.; Friedman, N.; Gnirke, A.; Regev, A. *Nature methods* **2010**, *7*, 709.

(51) Trapnell, C.; Pachter, L.; Salzberg, S. L. *Bioinformatics* **2009**, *25*, 1105.

Chapter 4: Chemoproteomic Discovery of Cysteine Containing Human sORFs

This chapter was adapted from:

Schwaid, A.G.*; Shannon, D.A.*; Ma, J.; Slavoff, S.A.; Levin, J.Z.; Weerapana, E.; Saghatelian, A. Chemoproteomic Discovery of Cysteine Containing human sORFs. Submitted

*authors contributed equally

4.1 Introduction

Our discovery of the prevalence of sORF-encoded peptides (SEPs) highlighted the extent to which the proteome is still unknown[1]. Our peptidomics based pipeline made major strides in identifying SEPs[1]. However, ribosome profiling and computational approaches suggested there were even more SEPs that our approach was failing to detect[2,3]. This suggested that different methods could be used to identify an even larger number of SEPs.

When analyzing complex samples, mass spectrometry is stochastically limited. This biases discovery towards the most abundant species in a sample. Therefore, it is plausible that a large number of SEPs could be missed by our mass spectrometry approach due to sample complexity. Therefore, in order to further enrich SEPs from cell lysate, and in order to access a different pool of SEPs we developed a chemoproteomics approach.

Here, we apply a cysteine affinity enrichment approach to identify novel cysteine containing SEPs (ccSEPs). Cysteine is the most reactive amino acid making it an ideal target for an affinity probe[4]. Additionally, 92% of proteins are estimated to contain at least one cysteine implying that the many SEPs should have a cysteine available for labeling[5]. Furthermore, cysteine reactivity is governed by secondary structure and local environment, suggesting that enriching SEPs with highly reactive cysteines will likely favor the discovery of SEPs with distinct secondary structures[4]. Most importantly, by using a different strategy to enrich the peptidome, we anticipate the discovery of novel ccSEPs.

4.2 Isolation of Cysteine Containing SEPs

Our strategy began with isolating the peptidome from K562 by lysis of these cells followed by a molecular weight cutoff (MWCO) filter to remove large proteins (Figure 1). We incubated the peptidome with a previously described iodoacetamide-alkyne (IA-alkyne) probe that reacts with the sulfhydryl side chain of cysteine to form a covalent bond to the peptide[1]. After cysteine capture by IA-alkyne, the probe is conjugated to a biotin-labeled tobacco etch virus (TEV) recognition peptide through copper-activated click chemistry (CuACC)[4,6,7]. Probe-labeled peptides are then separated from unlabeled peptides via streptavidin affinity chromatography to afford an enriched peptidome sample. On-bead trypsin digestion was performed, and unlabeled peptides were eluted and analyzed by ERLIC-LC-MS/MS[8]. The labeled peptides were then removed from the beads by the addition of TEV protease. This fraction was then analyzed by MudPIT-LC-MS/MS[9].

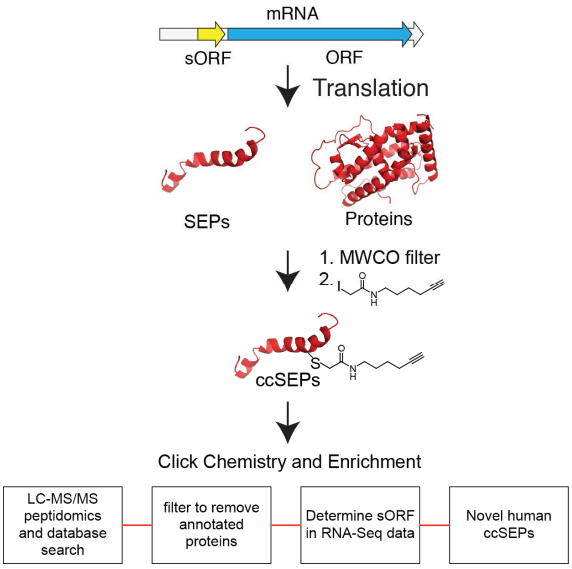


Figure 4.1: Workflow for identifying ccSEPs
The proteome and peptidome are separated by a MWCO filter and the peptidome fraction is carried forward to identify ccSEPS. Incubation of the peptidome with an iodoacetamide-alkyne (IA) probe leads to alkylation of cysteine-containing peptides including ccSEPs. Labeled peptides were then selectively enriched by conjugation to an azide-TEV-biotin tag using copper-activated click chemistry (CuACC) followed by affinity chromatography with streptavidin-coated beads. This sample is then analyzed by LC-MS/MS peptidomics and filtered to remove annotated proteins, which led to the identification of novel protein-generating sORFs that produce ccSEPs.

The data from this peptidomics analysis contains known as well as novel (i.e. non-annotated) peptides, including SEPs. In order to identify peptides originating from non-annotated RNAs, we created a custom database using K562 RNA-seq data[1], which contains information on the vast majority of mRNAs in K562 cells. Since these RNAs must be the source of any polypeptide produced we can include non-annotated genes in our peptidomics search by simply translating this database in three frames to generate a protein database that contains all possible peptide products. We then matched our peptide spectra against this RNA-seq database to reveal candidate SEPs. This approach yielded 175 hits that surpassed our preliminary Xcorr and deltaCN requirements[6]. After removing annotated peptides we were left with 109 candidate

SEPs. Our K562-RNAseq database was too large to perform a reverse database search directly. To overcome this, we constructed a forward and reversed database by appending our candidate SEPs to the human International Protein Index (IPI) database. We used this database to filter our candidate SEP spectra using a reversed database search, and only accepted peptides with a false discovery rate < 0.05. Subsequently, we validated that detected peptides could only originate from a sORF. Additionally, SEPs with more than 2 missed cleavages were removed along with SEPs detected from peptides fewer than 7 amino acids in length. Furthermore spectra were visually inspected to ensure good sequence coverage and confirm that peptides detected from the TEV fraction contained an IA-modified cysteine reside. After this, we were left with 17 novel human ccSEPs (Table 1).

Table 4.1: Detected peptides and the start codon and length (AUG or near cognate to stop) of their corresponding SEPs.

Detected Peptide[a]	Start codon	Length (aa)
C*GFFSYCSSESVSCSTS	ATC	34
STSLYCHSTILC*	AAG	24
TC*DGNSNEGGGTR	AAG	19
NFPLASSPERC*FFVPK	AAG	48
VEKLELLYIAGGNVNWYSPC*	GTG	22
YPAC*SPSPALI	CTG	29
GRGCC*RGFSAVGQGPSST	ATG	84
CPSINFQHFCHFVLCAFPIHC*	CTG	35
TC*TIPVPAGGRPR	CTG	32
IC*DIKGLIDNV	TTG	41
TSPADAVC*PGLGRDLCGSSRCCLRP	ATG	79
RGPGEAGMSWEEAGGLAPHLLC*CR	GTG	86
QIVLGGC*GEMV	alternate	16
GASFSEDGC*LLVG	CTG	37
GSSDIISVPC*	ATG	40
SSMPLIC*FLILEGLGR	ATG	29
CHFKIQLKGLLDLNTHT	ATG	97

[a]asterisk denotes probe labeled cysteine.

4.3 Validation of Cysteine SEP Labeling

To verify that our labeling and enrichment is specific to cysteine-containing SEPs, we performed an in vitro assay in cell lysates. We first synthesized TCT-SEP (named for the detected peptide; Table 1) by solid phase peptide synthesis, along with a mutant of this TCT-SEP where the cysteine is replaced by a serine, TST-SEP. We incubated TCT-SEP in K562 cell lysates and then added the IA-alkyne probe. After labeling, the lysate was mixed with a fluorescent azide in the presence of copper (II) sulfate and TCEP to promote CuACC. This fluorescently labeled lysate was then resolved on an SDS page gel to assess labeling of the TCT-SEP. Labeling of TCT-SEP was specific and robust and could be easily observed within total K562 lysate (Figure 2B). The control TST-SEP was not labeled when probe-treated alone or in K562 lysate demonstrating that labeling is occurring on the cysteine residue. Additionally, in cases where a detected peptide contained multiple cysteines, the labeled cysteine could always be determined by MS/MS fragmentation along the peptide backbone (Figure 3A).

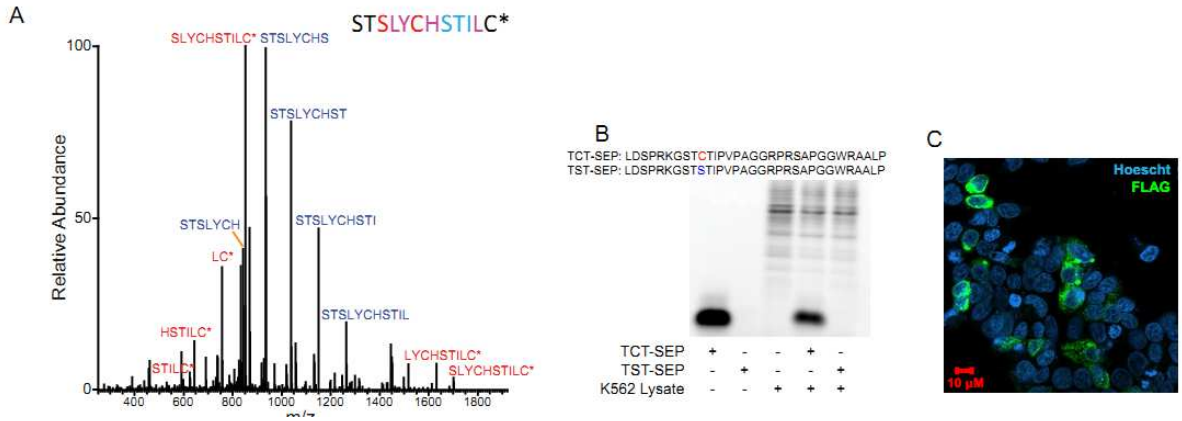


Figure 4.2: Validation of site of labeling and cellular expression of newly discovered ccSEPs. (A) In the case of ccSEPs with multiple cysteines, examination of the tandem MS spectra reveals the site of labeling. In this case, STS-ccSEP labels at the C terminal cysteine. Red indicates fragments detected by y ions, blue indicates fragments detected by b ions, and purple indicates fragments detected by both. (B) We tested labeling of one of the ccSEPs in a complex mixture by spiking the purified ccSEP into lysate and then performing a labeling reaction with rhodamine azide. If the ccSEP reacted it would fluorescently labeled. Mutation of the cysteine on the ccSEP to a serine abrogates labeling. (C) A C-terminal Flag tag appended to the sORF coding for TSP-ccSEP validated that this sORF does indeed produce protein. Staining of the protein product with an anti-Flag antibody confirmed expression and cellular stability of the ccSEP.

To validate the production of ccSEPs from their endogenous RNA, we transfected cells with a vector containing the sORF TSP-ccSEP, which is found on the same transcript as MRS2L. This construct contained the entire endogenous 5'UTR, which includes the sORF, and a FLAG tag was appended to the sORF to enable easy detection of protein production. Stable ccSEP expression was then observed by immunofluorescence using an anti-FLAG antibody (green) (Fig 3C.). This sORF was not annotated previously, thereby highlighting the ability of this workflow to discover novel protein-coding genes. More generally, this affinity strategy successfully identified a new pool of SEPs with characteristic hallmarks of this emerging class of peptides1.

4.4 Novel ccSEPs

An overview of these newly identified ccSEPs revealed many similarities with previously identified SEPs. First, the length of their sORFs ranged between 16 and 97 codons (Figure 2A). Second, these SEPs had both AUG start codons or non-canonical near cognate start codons (Figure 2B), similar to previously discovered SEPs. Moreover, SEPs could be found in the 3'UTR, frameshifted within known genes or within the 5' UTR, in non-annotated RNAs, or in antisense transcripts (Supporting information). All SEPs identified were fewer than 100 amino acids in length, and measuring sORF length as the stop codon to stop codon distance yielded a similar result. These identified SEPs are very small relative to the average length of a human protein, which is 335 amino acids[10]. The small size of these SEPs contribute to the difficulties associated with computationally predicting the presence of these peptides.

While specific functions for these SEPs will require additional downstream experiments, we wanted to see if we could gain some insights about those SEPs that may be most useful to investigate moving forward. Sequence conservation is an important and well-documented signifier of biological function[11]. We examined the conservation of our SEPs in several species by aligning in silico translated RNA. Of the SEPs we discovered over one third (6/17) are conserved amongst several species alluding to their potential function. Notably, the labeled cysteine residue does not vary between species, implying this residue may be important for the SEP's biological function (Figure 3C). Additionally, certain biologically important post translational modifications, such as protein S-nitrosylation, occur at, and can be regulated by, redox active cysteines[12]. The conservation of ccSEPs, particularly at redox active cysteine and surrounding residues, raises the possibility that ccSEPs could have a role in

regulating the cellular redox state. The conservation of these SEPs makes them good leads for further functional characterization, and demonstrates that this platform allows for the identification of peptides that are of significant biological interest.

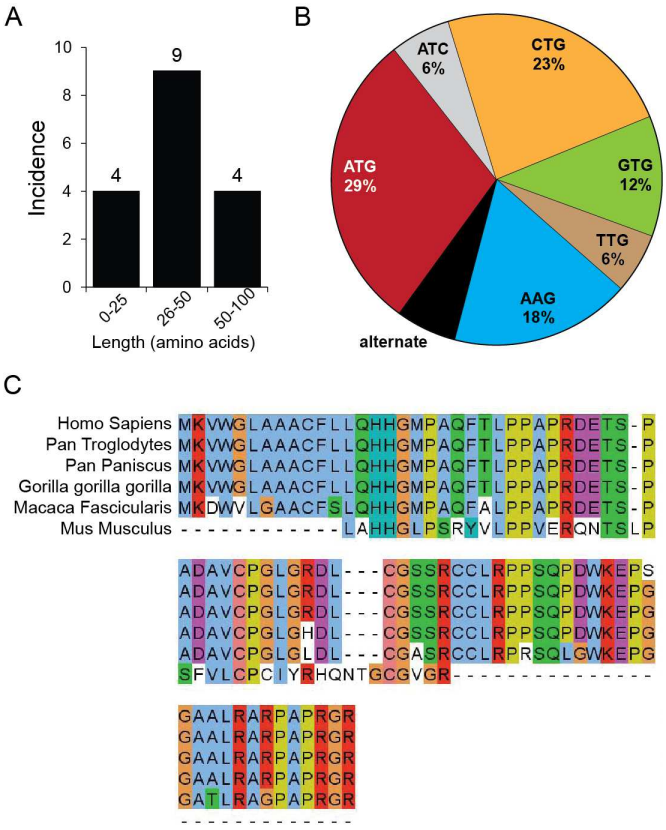


Figure 4.3: ccSEP overview. (A) Distribution of ccSEPs by their length in amino acids. SEP length was determined using the distance from an upstream in frame AUG start codon to a downstream in frame stop codon, or, when no inframe AUG was present, a near cognate start codon or stop codon was used instead. (B) While AUG is the predominant start codon for the production of ccSEPs, near congnate start codons (i.e. one base different from AUG) are also common. (C) TSP-SEP is strongly conserved amongst several species of primates suggesting this SEP may be functional.

4.5 Conclusion

In summary, we have utilized a chemoproteomics approach to identify new human ccSEPs. These results demonstrate the value of chemoproteomics to promote the discovery of additional sORFs. In this case, we identified seventeen novel ccSEPs indicating the presence of even more of these molecules than had been predicted, and representing a 16% increase in the number of known SEPs. Moreover, conservation indicates that some of these ccSEPs may be functional. The struggle to identify the whole range of SEPs in human cells as well as their functional role remains a key question in biology. The development of mass spectrometry methods focused on the identification of SEPs, such as chemoproteomic approaches, is a critical step towards answering these questions.

4.6 Methods

Cell culture:

Cells were grown at 37°C under an atmosphere of 5% CO_2. K562 cells were grown in RPMI 1640 medium with 10% FBS, penicillin and streptomycin. Cells were maintained between 1-10 x 10^5 cells/ml. HEK293T cells were grown in DMEM with 10% FBS, penicillin and streptomycin.

Isolation of polypeptides:

1x10^9 K562 cells were washed with ice cold PBS 3 times. Cells were subsequently suspended in lysis buffer with 0.1M ammonium acetate, 0.5M NaCl, diprotin A (1 ug/mL), antipain (1 ug/mL), leupeptin (1ug/mL), chymotrypsin (1 ug/mL) at pH 3.6 on ice. Cells were sonicated on ice for 20 bursts with output level 2 using 30% duty cycle (Branson Sonifier 250). Samples were then centrifuged at 3,000g for 10 min. Supernatant was collected and centrifuged through a 30kD molecular weight cutoff filter at 20,000g for one hour (Modified PES, Centrifugal Filter, VWR). Filtrate was dialyzed into PBS, and polypeptide concentration was measured by BCA assay. Cysteine containing SEPs were then enriched from this polypeptide sample for MudPIT-LC-MS/MS analysis as described below.

MudPIT-LC-MS/MS analysis:

Probe-labeling, click chemistry, and streptavidin enrichment:

Polypeptide samples were probe labeled with IA-alkyne (100 µM) for one hour at room temperature. Probe-labeled samples were subjected to click chemistry. Biotin-TEV-azide (200 µM), TCEP (1 mM, 50X fresh stock in water), ligand (100 µM, 17X stock in DMSO:t-Butanol 1:4), and copper(II) sulfate (1 mM, 50X stock in water) were added to the protein.

Samples were allowed to react at room temperature for 1 hour. Tubes were centrifuged (10 mins, 4°C) to pellet the precipitated proteins. The pellets were resuspended in cold MeOH by sonication. Centrifugation was followed by a second MeOH wash, after which the pellet was solubilized in PBS containing 1.2% SDS via sonication and heating (5 min, 80°C). The SDS-solubilized, probe-labeled proteome samples were diluted with PBS (5 mL) for a final SDS concentration of 0.2%. The solutions were incubated with 100 µL streptavidin-agarose beads (Thermo Scientific) at 4°C for 16 hrs. The solutions were then incubated at room temperature for 2.5 hrs. The beads were washed with 0.2% SDS/PBS (5 mL), PBS (3 x 5 mL), and water (3 x 5 mL). The beads were pelleted by centrifugation (1300 x g, 2 min) between washes.

On-bead trypsin digestion:

The washed beads were suspended in 6 M urea/PBS (500 µL) and 10 mM dithiothreitol (from 20X stock in water) and placed in a 65°C heat block for 15 mins. Iodoacetamide (20 mM, from 50X stock in water) was then added and the samples were placed in the dark and allowed to react at room temperature for 30 mins. Following reduction and alkylation, the beads were pelleted by centrifugation (1300 x g, 2 min) and resuspended in 150 µL of 2 M urea/PBS, 1 mM $CaCl_2$ (100X stock in water), and trypsin (2 µg). The digestion was allowed to proceed overnight at 37°C. The digestion was separated from the beads using a Micro Bio-Spin column (BioRad). The beads were washed with PBS (3 x 500 µL) and water (3 x 500 µL) to remove trypic peptides and urea.

On-bead TEV digestion:

The washed beads were resuspended in 150 µl of TEV digest buffer with AcTEV Protease (5 µl; Invitrogen) for 12 hr at 29°C with mild agitation. The eluted peptides were separated

from the beads using a Micro Bio-Spin column and the beads were washed twice with 75 µl water, and washes were combined with eluted samples. Formic acid (15 µl) was added to the samples, which were stored at −20°C until mass spectrometry analysis.

Offline electrostatic repulsion-hydrophilic interaction chromatography (ERLIC) fractionation of peptides:

ERLIC fractionation was performed offline prior to LC-MS/MS analysis using a PolyWax LP column (200 mm x 2.1 mm, 5µm, 300 Å, PolyLC Inc) connected to an Agilent Technologies 1200 Series HPLC equipped with a degasser and automatic fraction collector. Runs were performed with a flow rate of 0.3 ml/min. A 70 minute gradient beginning with 0.1% acetic acid in 90% acetonitrile and ending with 0.1% formic acid in 30% acetonitrile was used, and eluant was collected in 4 fractions. Fractions were evaporated to dryness before analysis by LC-MS/MS. Samples fractionated by ERLIC were not fractionated by SCX, and were loaded directly onto a C18 column for analysis.

Liquid chromatography-mass spectrometry (LC-MS) analysis:

LC-MS analysis was performed on an LTQ Orbitrap Discovery mass spectrometer (ThermoFisher) coupled to an Agilent 1200 series HPLC. Digests were pressure loaded onto a 250 µm fused silica desalting column packed with 4 cm of Aqua C18 reverse phase resin (Phenomenex). The peptides were eluted onto a biphasic column (100 µm fused silica with a 5 µm tip, packed with 10 cm C18 and 3 cm Partisphere strong cation exchange resin (SCX, Whatman)). Using a gradient 5-100% Buffer B in Buffer A (Buffer A: 95% water, 5% acetonitrile, 0.1% formic acid; Buffer B: 20% water, 80% acetonitrile, 0.1% formic acid). The peptides were eluted from the SCX onto the C18 resin and into the mass spectrometer following the four salt steps outlined in Weerapana et al[6]. The flow

rate through the column was set to ~0.25 µL/min and the spray voltage was set to 2.75 kV. One full MS scan (400-1800 MW) was followed by 18 data dependent scans of the n^{th} most intense ions with dynamic exclusion enabled.

MS data analysis:

The generated tandem MS data was searched using the SEQUEST algorithm against the databases listed in the main text. A static modification of +57 on Cys was specified to account for iodoacetamide alkylation, and a differential modification of 464.28596 was specified on Cys, corresponding to the IA-alkyne probe conjugated to the cleaved Biotin-TEV-azide tag. The SEQUEST output files generated from the digests were filtered using DTASelect 2.0. Samples with an XCorr score above 1.8 (+1), 2.5 (+2), 3.5 (+3) and deltaCN score above .08 were accepted.

Hits were then subjected to an iterative reverse database search. A reverse database was constructed by appending sORF encoded peptide sequences, which coded for unannotated detected peptides to the human IPI database. This database was reversed, and detected peptides were re-searched against the forward and reversed appended human IPI database. A 5% FDR threshold was set.

Peptide hits were then searched against the human IPI database using a string matching algorithm and matches were removed. Remaining hits were searched against the nonredundant protein database using Basic Local Alignment Search Tool (BLAST) and any peptides that matched known proteins were removed. All detected peptides consisted of 7 or more amino acids.

Spectra of the remaining peptides were manually validated to ensure a precursor mass error of < 10 ppm. Spectra also contained at least 5 sequential b or y ions, and

no more than 2 missed cleavages. In the case of peptides identified from the biotin eluted fraction, all peptides were labeled with an iodoacetamide probe at a cysteine residue.

SEP length was calculated using the length from the first AUG or near cognate start codon upstream of the detected peptide to the first stop codon downstream of the detected peptide. SEP length could also be calculated using the length from the first AUG or stop codon upstream of the detected peptide to the first downstream stop codon

4.7 References

(1) Slavoff, S. A.; Mitchell, A. J.; Schwaid, A. G.; Cabili, M.; Ma, J.; Levin, J. Z.; Budnik, B.; Rinn, J. L.; Saghatelian, A. *Nature Chemical Biology* **2012**, *9*, 59.

(2) Ingolia, N. T.; Lareau, L. F.; Weissman, J. S. *Cell* **2011**, *147*, 789.

(3) Stern-ginossar, N.; Weisburd, B.; Michalski, A.; Le, V. T. K.; Hein, M. Y.; Huang, S.; Ma, M.; Shen, B.; Qian, S.; Hengel, H.; Mann, M.; Ingolia, N. T.; Weissman, J. S. *Science* **2013**, *338*, 1088.

(4) Weerapana, E.; Wang, C.; Simon, G.; Richter, F.; Khare, S.; Dillon, M. B. D.; Bachovchin, D. A.; Mowen, K.; Baker, D.; Cravatt, B. F. *Nature* **2010**, *468*, 790.

(5) Miseta, A.; Csutora, P. *Molecular Biology and Evolution* **2000**, *17*, 1232.

(6) Weerapana, E.; Speers, A. E.; Cravatt, B. F. *Nature protocols* **2007**, *2*, 1414.

(7) Wu, P.; Feldman, A. K.; Nugent, A. K.; Hawker, C. J.; Scheel, A.; Voit, B.; Pyun, J.; Fréchet, J. M. J.; Sharpless, B. K.; Fokin, V. V. *Angew. Chem. Int. Ed* **2004**, *43*, 3928.

(8) Alpert, A. *Anal. Chem.* **2008**, *80*, 62.

(9) Washburn, M.; Wolters, D.; Yates, J. *Nature biotechnology* **2001**, *19*, 242.

(10) Ota, T.; Suzuki, Y.; Nishikawa, T.; Otsuki, T.; Sugiyama, T.; Irie, R.; Wakamatsu, A.; Hayashi, K.; Sato, H.; Nagai, K.; al., e. *Nature Genetics* **2004**, *36*, 40.

(11) Ponjavic, J.; Ponting, C.; Lunter, G. *Genome Res* **2007**, *17*, 556.

(12) Hess, D. T.; Matsumoto, A.; Kim, S.; Marshall, H. E.; Stamler, J. S. *Nature Reviews Molecullar Cell Biology* **2005**, *6*, 150.

Chapter 5: Functional Characterization of sORF-Encoded Peptides

Aravind Subramanian, Willis Read-Button, and Ted Natoli measured the L1000 data. Rajiv Narayan processed and Z-scored the L1000 data. I prepared samples for L1000 measurement, performed downstream bioinformatic analysis, and performed all other experiments.

5.1 Introduction

Mass spectrometry and ribosome profiling methods have successfully been used to identify sORF-encoding polypeptides (SEPs)[1,2]. These studies have unequivocally demonstrated widespread translation from non-annotated sORFs, and therefore SEPs represent the cells protein 'dark matter'. Additionally, evidence of translation from non-AUG start codons implies the proteome is larger and more complicated than imagined[1,3]. In light of these discoveries, a major challenge moving forward is the determination of the functions of these SEPs, if any, in cells and/or tissues.

A handful of SEPs have been found to be functional in other species and one has been found in humans. The eleven amino acid peptide encoded by the *Tal* gene is necessary for proper morphogenesis in *Drosophila*[4]. *Tal* is a polycistronic mRNA that encodes three 11 and one 32 amino acid long peptides[5]. These peptides were found to interact with filamentous actin and are involved in denticle formation. In the absence of the *tal* gene, or when the sORFs are frameshifted, apical cuticular structures are completely eliminated. These examples highlight the importance of SEPs in fly morphogensis, and also highlight the misannotation of key genes as non-coding RNAs.

A bioactive SEP has also been found in human cells. Humanin, a 24-amino acid peptide, was shown to inhibit neuronal cell death induced by familial Alzheimer's disease mutant genes and amyloid-β (Aβ)[6,7]. Using yeast two hybrid assays, it was determined that humanin mediated these functions by interacting with insulin-like growth factor-binding protein 3[8], and cell biology has determined that humanin operates by inhibiting the function of the pro-apoptotic BAX. Of course, these SEPs were discovered through functional screens and therefore it still remains to be determined

whether any of the SEPs discovered through LC-MS profiling, or ribosome profiling for that matter, are actually functional.

It has been suggested that SEPs are involved in *de novo* gene birth, and are in fact proto-genes[9]. This theory suggests that translation of sORFs to SEPs is a way of sampling non-genic regions of the genome for peptides that confer an adaptive advantage. Presumably, such peptides would be selected for and eventually evolve into more strongly conserved genes. This is a compelling hypothesis, but it is possible that sORFs could prove advantageous without evolving into longer genes. This suggests the possibility that sORFs themselves could have function, and emphasizes the need to develop systematic tools to understand SEP function in order to fully understand the biology SEPs may regulate.

The need to systematically understand whether or not a class of biomolecules is functional and what its function may be is not a new one. Guttman et al. faced this challenge when confronting the discovery of the existence of thousands of long intergenic noncoding RNAs (lincRNAs)[10]. In order to determine whether or not these molecules are functional and to assign them putative functions, they developed a strategy dependent on measuring changes in gene expression. In particular, they used DNA microarrays to analyze lincRNA levels in different cellular or tissue contexts. For instance, Guttman and coworkers looked at lincRNA levels over a time course of embryonic tissue development, in different adult tissues (brain, lung, ovary, and testis), and in several cell lines (mouse ESCs, NPC, MEF, and MLF). They then categorized expression levels in these tissue and cell types, and correlated lincRNA levels to these categories. This allowed them to assign lincRNAs into basic Gene Ontology

categories[11]. Additionally, they could validate some of these putative assignments. For example, they treated *p53*$^{+/+}$ and *p53*$^{-/-}$ MEFs with a DNA damaging agent and observed that lincRNAs they had assigned to the category of "p53-mediated damage response" were significantly elevated in the *p53*$^{+/+}$ but not *p53*$^{-/-}$ MEFs. This strategy successfully demonstrated lincRNAs were functional and hinted at their possible cellular functions. On its face, a similar strategy could be used to understand SEPs. In order to systemically screen for SEP bioactivity, and gain insights into SEP function we have developed a causative transcriptomics-based pipeline to characterize SEPs (Figure 5.1).

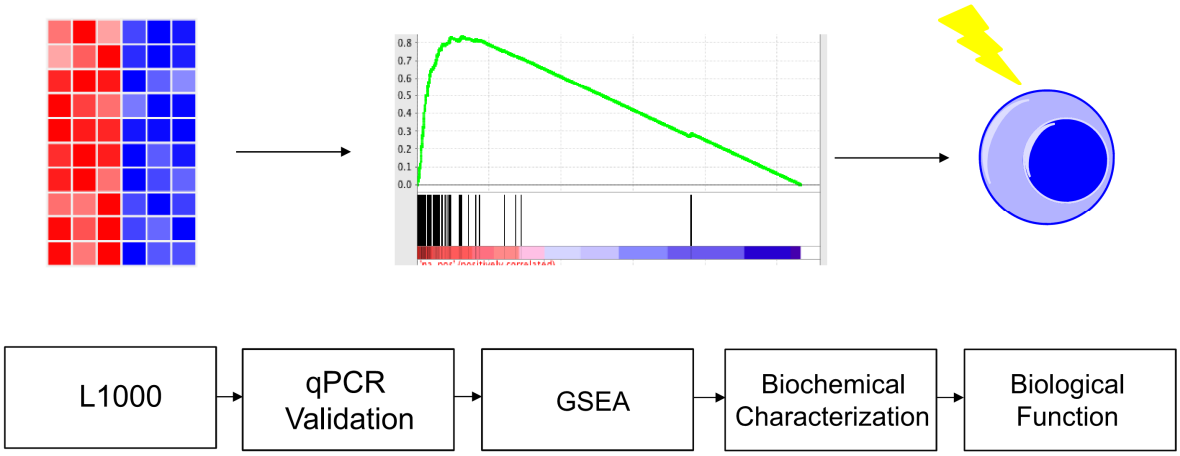


Figure 5.1: Platform for functionally characterizing SEPs. In order to functionally characterize SEPs we screened for changes in gene expression upon SEP overexpression. The most interesting cases could then be validated by qPCR. Changes in gene expression could be mapped to pathways using gene set enrichment analysis (GSEA). This yielded a putative cellular role for the SEPs, which could be further characterized biochemically and these changes could be ascribed to a biological function.

5.2 SEPs alter gene expression

Changes in gene expression upon knockdown or overexpression of a protein are one possible signifier of function. Since SEPs are frequently bicistronic[1] it is often impossible to knockdown the SEP without knocking down an annotated protein coding gene. This would greatly complicate any interpretation of changes in gene expression. In order to simplify data interpretation, and ensure changes in gene expression are a result of changes in SEP expression we opted to increase cellular SEP levels in order to determine whether or not SEPs effected gene expression. Moreover, since all known SEPs are thought to be intracellular and SEPs are too large to pass through the cell membrane we chose to recombinantly overexpress SEPs within the cell.

We transiently transfected a vector coding for flag tagged SEPs in HEK293T cells. The presence of a flag tag allowed us to confirm translation and expression of SEPs via immunofluorescence. As a control, HEK293T cells were transfected with an empty vector. Afterwards, cells were lysed, mRNA was harvested, and gene expression levels were measured with a streamlined microarray technology called L1000.

L1000 is a method for quickly and cheaply measuring the transcript levels of 1,000 genes, and computationally inferring the transcript levels of an additional 21,000 genes[12]. Using this method, mRNA is reverse transcribed to cDNA. cDNA is then annealed and ligated to upstream and downstream probes which contain gene specific sequences, an oligonucleotide “barcode”, and upstream and downstream primer sites. The ligated product is then amplified, and the 5’ end is conjugated to biotin. Fluorescent microspheres conjugated to oligonucleotides are annealed to the DNA “barcode”, and amplicons are immobilized. The fluorescence of the annealed microspheres is

measured, and the fluorescence intensity is proportional to the abundance of transcripts of the gene complementary to the gene specific probes. This approach can be simultaneously applied to 1000 transcripts to directly measure their abundance. Subsequently, using inferential models developed at the Broad Institute the expression levels of an additional 21,000 genes can be estimated. L1000 is a critical technique for measuring expression levels in response to SEP treatment. Although DNA microarrays would provide similar data, this method lacks the throughput and affordability of L1000. Using this technique allowed for the gene expression levels of many SEP treated samples to be measured.

Using this approach we measured changes in gene expression after treatment with 20 different SEPs. Of these SEPs, 17 resulted in significant changes in gene expression and three did not though they were expressed. Thus, not all SEPs regulate transcription. Gene expression changes were determined to be significant if they surpassed a rigorous threshold of repeatability and signal strength. Namely, gene expression signatures had to have a pairwise spearman correlation coefficient of ≥ 0.2 within the 75^{th} percentile, and a signal strength score of ≥ 4 where this number was computed as the difference in means between the top and bottom 50 most differentially expressed genes.

To illustrate the strength of these gene expression changes, cells were also treated with a variety of small molecule signal calibrators (Figure 5.2). These calibrators included wortmanin, a kinase inhibitor, geldanamycin, an HSP90 inhibitor, and HDAC inhibitors Vornistat and Trichostatin A[13-16]. Most SEPs induced gene expression changes similar in magnitude to treatment with wortmanin, which inhibits

phosphotidylinositide 3-kinases (PI3Ks)[13]. One SEP induced gene expression changes similar in magnitude to inhibition of HSP90 with gelandamycin, which leads to degradation of p53[14]. Two SEPs induced gene expression changes even larger than this, although they did not induce changes in the same range as treatment with HDAC inhibitors[15,16]. These changes in gene expression indicate SEP expression has a strong effect on gene expression, and demonstrates that the majority of SEPs are able to effect gene expression levels, which suggests SEPs may have important roles within the cell.

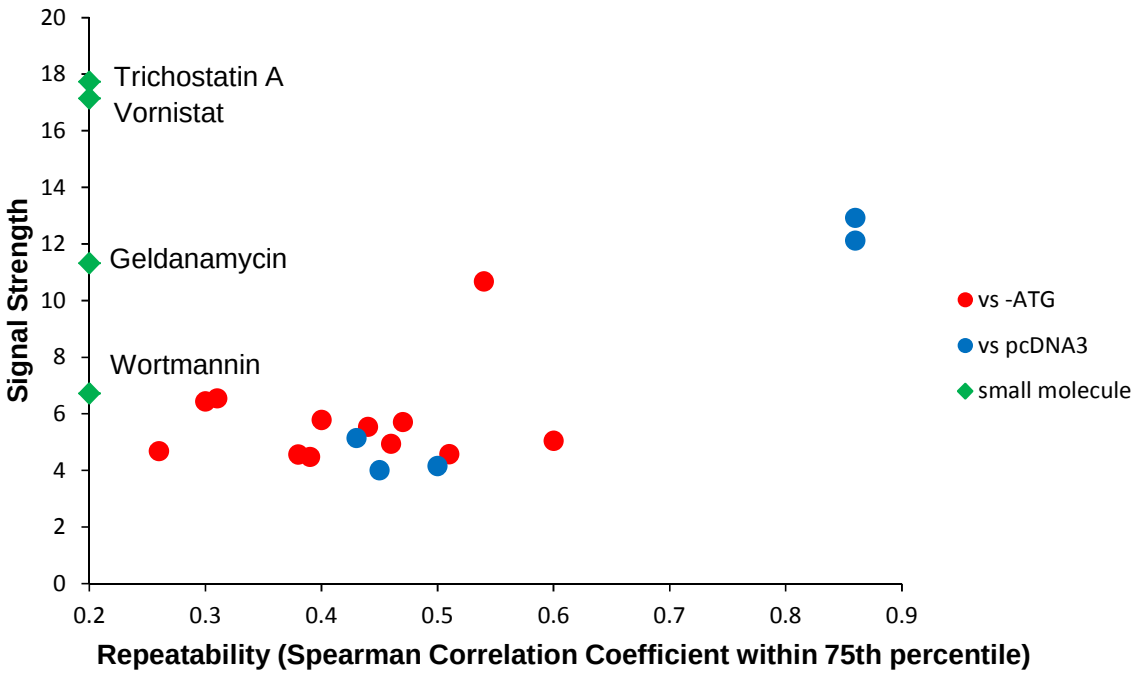


Figure 5.2: Signal strength and repeatability of gene expression changes induced by SEPs. Signal strength was calculated as the mean difference in expression of the top 50 and bottom 50 compared to the control. Repeatability is indicated by the Spearman Correlation Coefficient within the 75th percentile. Cells expressing SEPs were compared to cells transfected with vectors bearing the SEP sORF but lacking an initiation codon or to cells treated with pcDNA3. Additionally, cells were treated with trichostatin A, vornistat, HSP90, or wortmannin as signal strength calibrators.

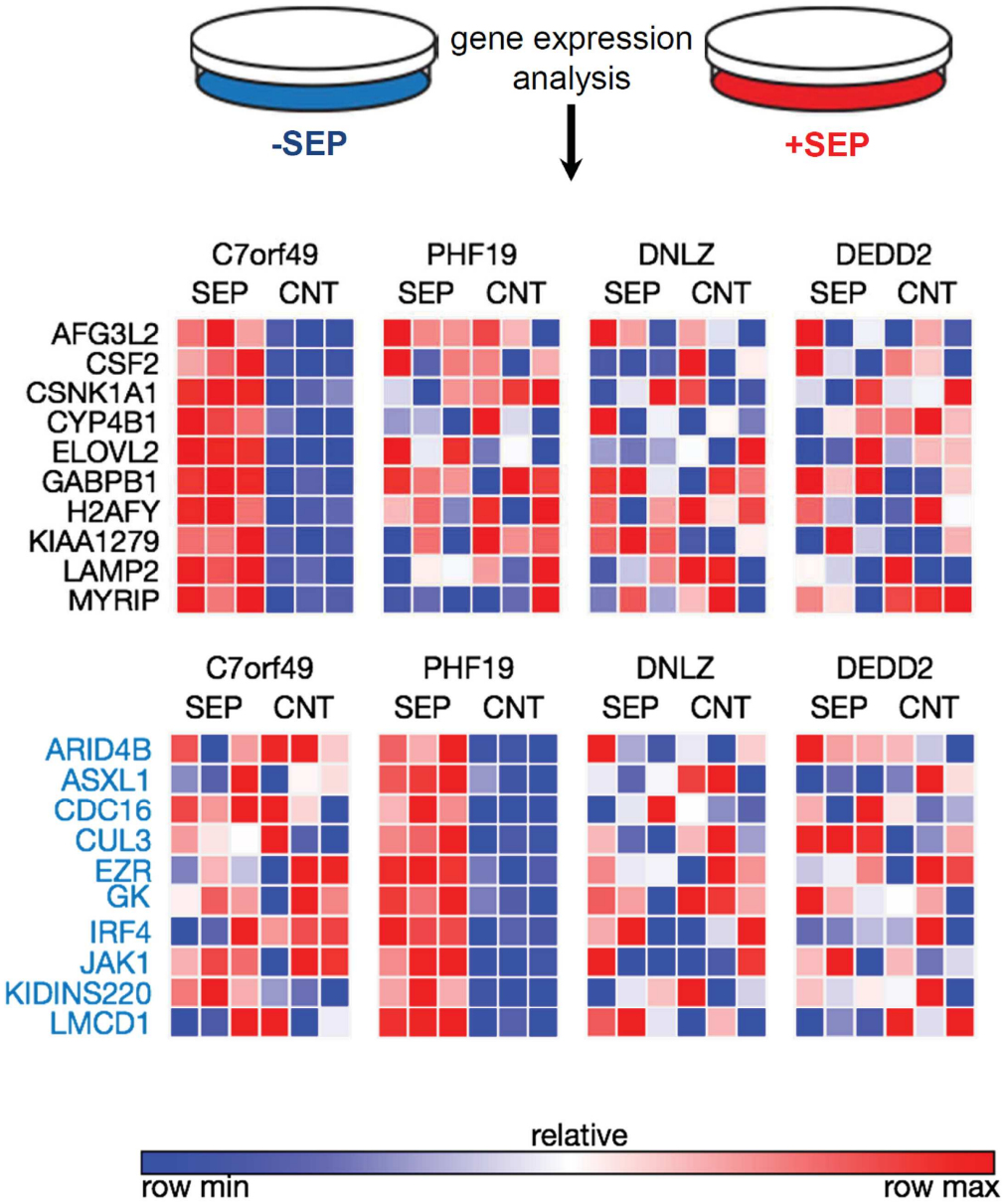


Figure 5.3: Marker analysis indicates SEPs induce specific gene expression changes. Analysis of the top 10 most highly induced genes upon SEP expression demonstrates C7orf49 regulates different genes than PHF19-SEP, DNLZ-SEP, or DEDD2-SEP. Likewise, the most highly induced genes upon PHF19 expression are not induced by C7orf49-SEP, PHF19-SEP, DNLZ-SEP, and DEDD2-SEP. This indicates gene expression changes are specific to the SEP polypeptide sequence.

Changes in gene expression upon SEP treatment were specific (Figure 5.3).

Genes whose expression levels were affected by one SEP were not affected by

another. Marker analysis, where the top ten most overexpressed and under expressed genes are analyzed, indicated no pattern in gene regulation between different SEPs. Moreover, gene set enrichment analysis (GSEA) performed between SEPs demonstrated no overlap (Figure 5.4)[17]. This indicates that SEP induced changes in gene expression are not a broad cellular response to overexpression of peptide and are rather an exclusive effect induced by the amino acid sequence of each particular SEP.

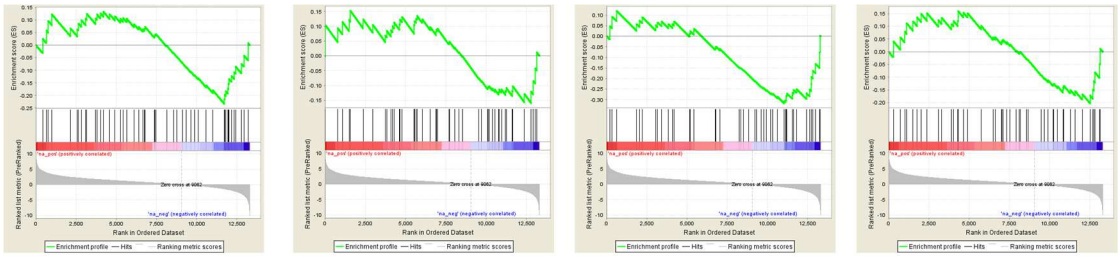


Fig 5.4: Gene set enrichment analysis indicates SEPs induce specific gene expression changes. Gene expression changes induced by EIF5-SEP were compared to gene sets comprising the top 50 most upregulated genes upon DEDD2-SEP, MCG16703-SEP, PHF19-SEP, or C19orf29SEP expression. None of them matched further supporting that SEPs induce different gene expression changes from one another.

Another possibility is that changes in gene expression are due to changes at the RNA level. Namely, that expression of the sORF RNA, but not the SEP itself resulted in regulation of gene expression. To eliminate this possibility, the start codon was eliminated from 14 out of 20 sORFs and a vector coding for this non translatable sORF was transfected in place of an empty vector. In the 14 cases analyzed ablation of the start codon was found to eliminate translation of the sORF. However, in 4 additional cases ablation of the putative start codon did not eliminate translation. In these cases it is possible a truncated SEP is produced from alternate start codons within the sORF. Notably, this is consistent with ribosome profiling data, which indicates a single sORF

can initiate with multiple start codons[2]. Comparison between vectors encoding SEPs and vectors encoding non-translatable SEPs demonstrated that changes in gene expression were dependent on SEP translation indicating gene expression is regulated at the peptide, not the RNA level (Figure 5.2). Together, these experiments demonstrate that SEPs regulate gene expression, and present the first evidence that SEPs, as a class of biomolecules, are bioactive.

5.3 SEPs can be assigned to putative cellular processes

Changes in gene expression can be related to biological phenomena through a variety of conceptual techniques including hierarchal clustering, marker analysis, and gene set enrichment analysis (GSEA) amongst others[17]. In order to interpret the biological significance of SEP induced changes in gene expression, we used GSEA to match changes in gene expression to putative pathways and processes. This method provides a robust and straight-forward route to biological characterization. GSEA allows one to compare changes in expression data to known grouping of genes that are up or downregulated in a particular process. These groupings are called gene sets, and can be relatively simple—such as a group of genes categorized in a Gene Ontology defined process—or can be more complicated such as gene changes observed upon cellular treatment with perturbogens. Using GSEA gene expression changes stimulated by SEPs could be analyzed.

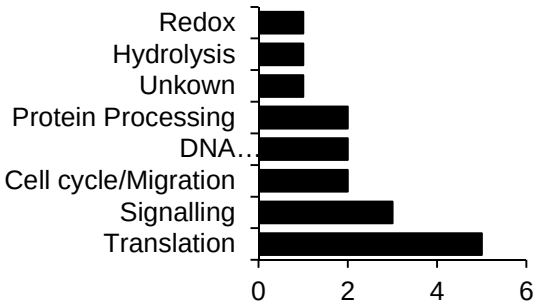


Figure 5.5: SEPs are involved in cellular pathways. SEP induced gene expression changes can be mapped to putative pathways by using gene set enrichment analysis to match expression profiles to Gene Ontology gene sets. Similar to an assortment of unrelated proteins, SEPs were involved in a variety of different pathways.

SEPs were mapped to basic cellular processes and pathways by means of GSEA against Gene Ontology gene sets (Figure 5.5). This analysis was successful at assigning putative pathways or cellular processes for 16/17 SEPs. The remaining SEP did not provoke gene expression changes that fell neatly within a Gene Ontology category. However, gene expression changes induced by this SEP could be mapped to gene expression changes induced by various perturbogens, or present in certain tissue states. Thus, although this SEP may not have a clear role in any single pathway it does regulate genes involved in various cellular processes.

The 16 SEPs that mapped distinctly to particular cellular processes fell into several different categories. In particular, SEPs were involved in protein processing, DNA binding and transcription, cell cycle and migration, signaling, translation, and pathways involved in hydrolysis or redox. Interestingly, SEPs did not appear to be involved in a single sphere of biology, but rather were involved with numerous unrelated pathways. This suggests that SEPs are involved in disparate processes. This implies

SEPs are important in a broad range of biological processes, as is the case for known proteins.

5.4 eIF5-SEP is involved in inflammation

Expression of eIF5-SEP resulted in a signal strength score of 12.9, indicating larger gene expression than any other SEP. Gene expression changes induced by eIF5-SEP were dramatic, and were substantially higher than gene expression changes induced by inhibiting PI3Ks which had a signal strength score of 6.7 or HSP90 which had a signal strength score of 11.3. We validated changes in several of the genes observed to change in the presence of eIF5-SEP L1000 by qPCR (Figure 5.6). The magnitude of these changes suggested eIF5-SEP may have an important role in biology and merited further study.

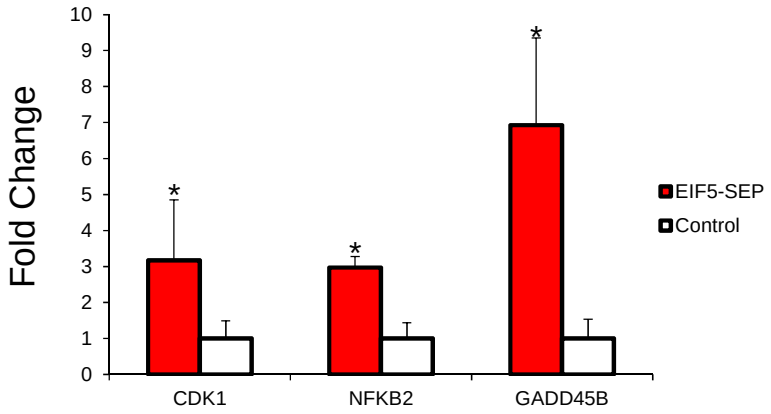


Figure 5.6: QPCR validates eIF5-SEP gene expression changes. We validated the expression of 3 genes indicated to be highly expressed in the L1000 data using qpcr. Statistics was performed with a Student's t-test, *, $p \leq 0.05$. One tailed t-test was used to confirm the one directional hypothesis developed from the L1000 data.

In order to understand at a detailed level what effect eIF5-SEP may have on cell biology we compared the gene expression profile of eIF5-SEP treated cells to a database of chemical and genetic perturbations gene sets. In these gene sets cells or

tissues are perturbed and changes in gene expression are measured. These changes are stored as a gene set. For instance, cells could be treated with a small molecule, or UV light, and changes in gene expression after treatment could be measured and stored. Additionally, gene expression in cells with abnormal genetic backgrounds—such as cancer cells—can be catalogued.

This analysis revealed that eIF5-SEP expression regulated genes also regulated by treatment with TNFα[18,19]. That is, genes overexpressed when cells are treated with TNFα were also overexpressed when cells overexpressed eIF5-SEP. This result indicates that, like TNFα, eIF5-SEP is involved in inflammation (Figure 5.7). Moreover, of the 19 other SEPs examined for changes in gene expression only one, TRIM41-SEPmapped to similar gene sets indicating these changes in gene expression are specific to eIF5-SEP. Notably, the genes eIF5-SEP activates are downstream of NF-κB, and NF- κB itself is overexpressed in the presence of eIF5-SEP (Figure 5.7).

Interestingly, when matched against Gene Ontology gene sets (which do not contain a NF- κB geneset) eIF5-SEP matched most strongly to genes involved in cellular redox activity. The redox state of the cell is an important determinant for NF-κB activity, and can effect NF- κB translocation into the nucleus and subsequent DNA binding[20]. In turn, NF- κB can influence the redox state of the cell by inducing transcription of ferritin heavy chain (FHC1) and superoxide dismutase 2 (SOD2)[21]. This suggests the possibility that eIF5-SEP is effecting NF-κB target gene expression by altering the redox state of the cell, although could also be interpreted as a change in the cellular redox state first induced by eIF5-SEP stimulation of NF-κB target gene expression. In contrast, TRIM41-SEP triggered genes involved in protein kinase

signaling suggesting that although stimulation with eIF5-SEP and TRIM41-SEP may have similar outcomes they operate through different mechanisms. Regardless of precisely how eIF5-SEP induces changes in NF- κB target gene expression, these changes provide a strong indication that eIF5-SEP has a role in inflammation through activation of the NF- κB pathway.

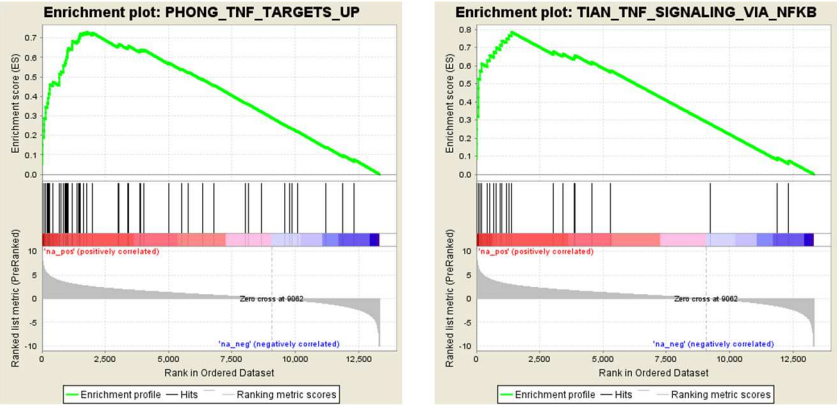


Figure 5.7 eIF5-SEP induces expression of pro-inflammatory genes. Gene expression changes induced by EIF5-SEP mapped to gene expression changes induced by TNFα. In particular, EIF5-SEP induced expression of NF-κB and NF-κB genes, which are involved in inflammation.

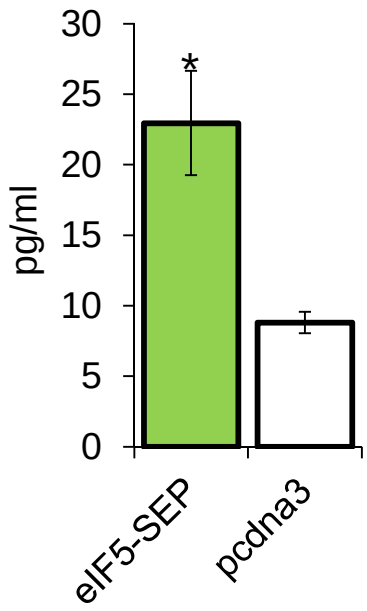


Figure 5.8: eIF5-SEP induces expression of IL-8. Media from cells transfected with eIF5-SEP had much higher levels of IL-8 than media from cells transfected with pcdna3 as measured by ELISA. IL-8 is a key mediator of inflammation supporting the role of EIF5-SEP in inflammation. Statistics was performed with a Student's t-test, *, $p \leq 0.05$.

Interleukins are a key intercellular regulator of inflammation[22]. Furthermore, certain interleukins are known to be upregulated upon treatment with TNFα, which stimulates NF-κB[23]. In order to see whether or not overexpression of eIF5-SEP was consistent with TNFα treatment at the protein level, and also whether or not eIF5-SEP treatment resulted in higher levels of proinflammatory proteins, we performed an ELISA to measure interleukin 8 levels (IL-8), and found that HEK293T cells transfected with eIF5-SEP secreted 2.6 fold more IL-8 as compared to control (Figure 5.8). IL-8 is a cytokine and chemotactic agent for lymphocytes and neutrophils[24]. Its expression also induces angiogenesis, and it has a well described role in acute inflammation[25,26]. Increase in IL-8 levels upon EIF5-SEP treatment is a strong indicator that EIF5-SEP is involved in inflammation.

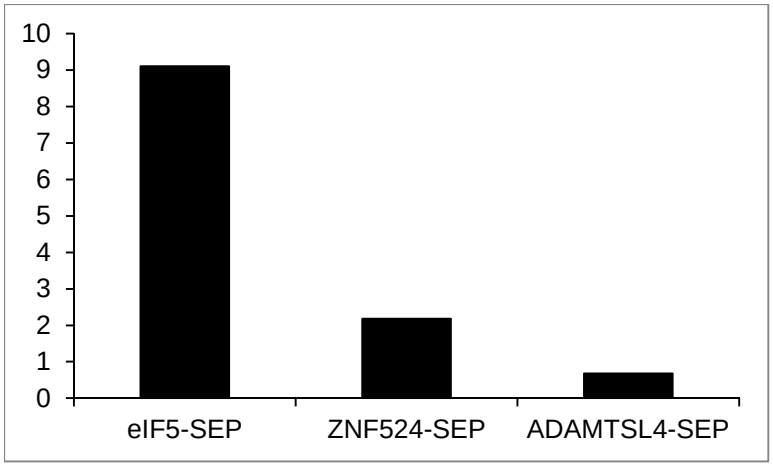


Figure 5.9: IL-8 RNA expression. eIF5-SEP, but not ZNF524-SEP, or ADAMTSL4-SEP induces expression of IL-8 RNA. This measurement is a Z-score derived from L1000 experiments.

Changes in IL-8 levels were consistent at the genetic level as well. L1000 results indicated highly upregulated transcription of IL-8 RNA upon eIF5-SEP stimulation (Figure 5.9). Moreover, stimulation with other SEPs did not affect IL-8 RNA levels. This indicates that pro-inflammatory effects of eIF5-SEP are particular to this SEP.

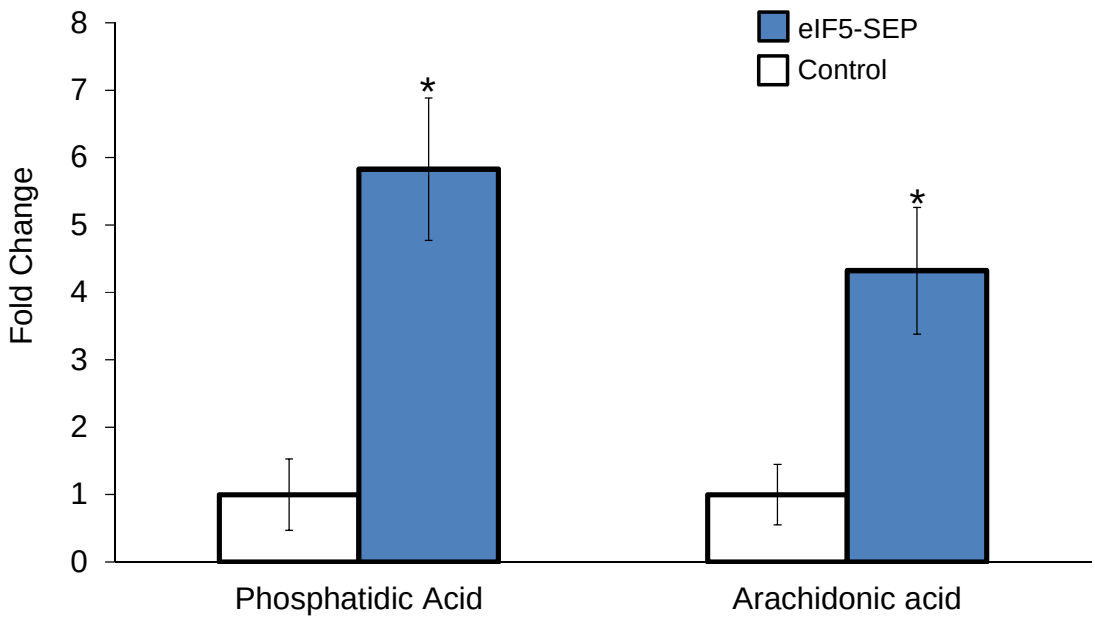


Figure 5.10: eIF5-SEP upregulates pro-inflammatory metabolites. Cellular lipid levels were analyzed upon eIF5-SEP transfection by LC-MS using a lipidomics approach. This approach demonstrated elevated levels of pro-inflammatory metabolites, phosphatidic acid, and arachidonic acid. The phosphatidic acid exact mass corresponds to 20:1/20:4 phosphatidic acid. Statistics was performed with a Student's t-test, *, $p \leq 0.05$.

Pro-inflammatory signaling can also be triggered by changes at the metabolite level. Global metabolite profiling of cells after transfection with vector coding for eIF5-SEP indicated upregulation of a phosphatidic acid with a mass corresponding to 20:4/20:1 phosphatidic acid and arachidonic acid (Figure 5.10). Arachidonic acid is a pro-inflammatory metabolite known to be upregulated upon TNFα treatment that can be metabolized to leukotrienes, prostaglandins and thromboxanes[27,28]. Phosphatidic acid is a precursor to arachidonic acid and can therefore be involved in inflammation as well. In the presence of cytosolic phospholipase A2 (PLA2), which is upregulated in cells treated with eIF5-SEP or TNFα, the 20:4 fatty acid of phosphatidic acid is hydrolyzed from the glycerol backbone resulting in the formation of arachidonic acid, and thus

contributes to proinflammatory signaling[29]. These results are consistent with a proinflammatory role for eIF5-SEP.

Expression of eIF5-SEP results in changes at the RNA, protein, and metabolite level that are all consistent with a pro-inflammatory role. Moreover, changes induced by eIF5-SEP mimic pro-inflammatory changes induced by TNFα. This evidence clearly points to eIF5-SEP as an important peptide in inflammation.

5.5 *eIF5-SEP* regulation

eIF5-SEP is located within the 5'UTR of another gene, *elongation initiation factor 5 (eIF5*). Previously, we have demonstrated that some SEPs are expressed bicistronically[1]. However in the case of bicistronic 5'UTR SEP expression, expression initiated with an alternate start codon, and leaky translation of the 5' ORF allowed for translation of the downstream gene. However, *eIF5-SEP* initiates with an AUG start codon, and, though frameshifted, overlaps with *eIF5*. This would seem to down-regulate or prevent translation of the downstream eIF5 gene[1,30,31].

eIF5 is a component of the ribosome, and after the 40S ribosome subunit has bound to a start codon, eIF5 mediates joining of the 40S and 60S ribosome subunits and hydrolysis of GTP bound to eIF2[32]. Thus, eIF5 is necessary for cell growth and viability[33,34]. Therefore, the organization of *eIF5-SEP and eIF5* posed the question of how both of these genes could be translated.

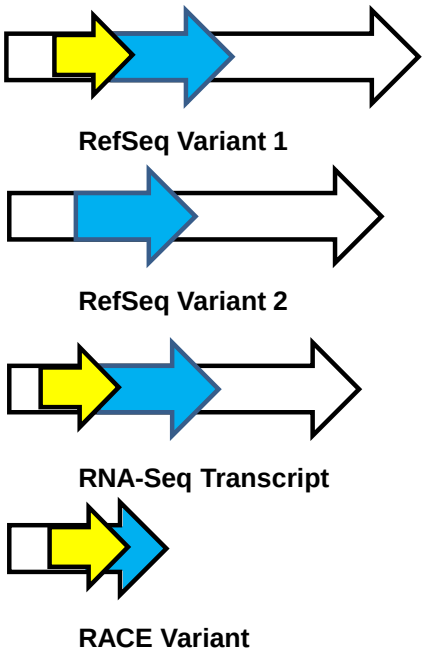


Figure 5.11: RACE PCR illustrates a mechanism of EIF5, EIF5-SEP translation. Rapid amplification of cDNA ends (RACE) using forward and reverse gene specific probes for EIF5-SEP detected an unidentified transcript, which contains the EIF5-SEP sORF and truncated EIF5. The 3' end of EIF5 is truncated and encodes only 161/431 codons of EIF5, although the entirety of EIF5-SEP is encoded. This provides a mechanism for how both EIF5 and EIF5-SEP can be translated even though their coding sequences overlap one another.

Annotated transcripts of *eIF5* indicate two transcript variants, one containing both *eIF5-SEP* and *eIF5*, and another transcript with a truncated 5' UTR that only encodes *eIF5* (Figure 5.11)[35,36]. The presence of this second transcript explains how *eIF5* is translated. However, it appeared *eIF5-SEP* could only be translated from the first transcript variant. This implied *eIF5-SEP* is coregulated with *eIF5*. This was surprising, since these two genes do not appear to share any common function. Therefore, in order to more fully understand how *eIF5-SEP* is translated we performed RACE-PCR to identify any unannotated transcript variants of this gene. Notably, we discovered a transcript with a truncated 3'UTR. This transcript encodes the *eIF5-SEP*, and does not code for the full length *eIF5* gene. Thus, *eIF5-SEP* and *eIF5* are both able to be

translated as they are encoded on separate transcripts. This data presents an alternate mechanism of regulation of 5' sORF translation in addition to bicistronic translation, and demonstrates a means by which eIF5-SEP can be expressed without coregulation of eIF5 at the translational level.

5.6 Conclusion

In this chapter we were able to develop a transcriptomics based pipeline, which revealed that SEPs are bioactive, and identified one SEP that has a significant role in inflammation. This approach allowed us to characterize a class of molecules with no *a priori* knowledge of their function, or whether indeed they had any function at all. Moreover, this approach allowed us to screen for molecules that induced the most significant cellular changes and allowed us to develop a lead for follow-up biological characterization. The success of this approach allowed us to make major strides in understanding the biological role of sORF peptides.

An unresolved question regarding SEPs biological function is their apparent coregulation with annotated proteins. Since many SEPs are located on the same transcript as annotated genes, and translated bicistronically many SEPs must be coregulated with known genes at least at the RNA level. As the study of SEPs moves forward, the purpose of this coregulation stands out as a tantalizing question. On one hand, encoding two ORFs in close proximity could merely be a way to incorporate the maximum amount of information in a minimal amount of space, or could arise by chance as functional sORFs evolve from proto-genes. Another explanation is that the two ORFs regulate the same process, or are expressed in response to the same stimulus. A key difficulty in answering this question is that it requires the complete functional

characterization of both SEP and protein. However, from our transcriptomics based approach it does not appear that this is likely, since the genetic signature induced by SEP stimulation does not overlap with the known functions of proximal genes. As the function of the proteome becomes more fully understood it will be exciting to see how the answer to this question unfolds.

Ultimately the fusion of mass spectrometry and transcriptomics based approaches has allowed for the discovery of a new class of functional biomolecules. Moreover, this approach has made inroads into understanding exactly how SEPs affect biology. These discoveries highlight that, although the Human Genome Project was completed a decade ago, the complexity and breadth of the proteome it encodes is still not fully known. The discovery and characterization of SEPs open new avenues to explore, and will allow us to understand biology at a deeper level.

5.6 Methods:

Cell Culture:

HEK293T cells were grown in DMEM with 10% FBS and penicllin/streptomycin. Cells were grown at 37°C with a 5% CO_2 atmosphere.

L1000 Transfection:

Cells were transfected at 70% confluency with 250ng plasmid DNA and .5uL lipofectamine 2000 in optimem. 10, 20, or 30 hours later cells were harvested. To harvest cells, media was aspirated and replaced with 100uL TCL buffer (qiagen) gently rocked for 30 minutes, then covered and frozen at -80°C until measurement.

L1000 Measurement:

The L1000 measurement was performed as reported in Peck et al.[12]

Bioinformatics:

Bioinformatic analysis was performed with the Gene-E software and Msigdb software freely available on the Broad's website (http://www.broadinstitute.org/gsea/msigdb/index.jsp) In order to categorize SEPs into putative cellular pathways GSEA was run against the c5 Gene Ontology data set. In order to assign more specific functions to SEPs GSEA was run against the c2 dataset. Both datasets are freely available at (http://www.broadinstitute.org/gsea/msigdb/index.jsp and http://www.broadinstitute.org/cancer/software/GENE-E/)

Heat maps were made in Gene-E software suite, where SEP signal strength is determined as a function of the signal strength of the well compared to the average signal strength of the plate as background. GSEAs were conducted using a preranked

list, ranked by Z-score. Z-score was determined by (). For gene set matches, only matches with an FDR≤.05 and p value≤.05. SEPs were assigned to putative pathways based upon the geneset with the highest normalized enrichment score from a GSEA versus the c5 Gene Ontology gene set. Positive phenotype matches were given priority to negative phenotype matches, since this simplifies downstream interpretation.

QPCR:

QPCR was conducted with the qiagen quantitect reverse transcription kit and qiagen quantitect SYBR green kit according to the manufacturer's instructions. Primers were:

GAPDH: GGCTCTCCAGAACATCATCCCTGC

GAPDH: GGGTGTCGCTGTTGAAGTCAGAGG

NFKB: CCTGGCAGGTCTACTGGAGG

NFKB: AAATAGGTGGGGACGCTGT

CDKN1A: GGAGGCGCCATGTCAGAACCGGCT

CDKN1A: GCCATTAGCGCATCACAGTCGCGGCTC

GADD45B: ACATGACGCTGGAAGAGCTCGTGGCG

ELISA:

IL-8 ELISA was performed using a kit from invitrogen according to the manufacturer's instructions

Metabolite Profiling:

10cm dishes were transfected with plasmid coding for EIF5-SEP or pcdna3. Twenty four hours later, cells were washed and harvested. Cells were extracted into 2:1:1 chloroform:methanol:media. The chloroform layer was dried under nitrogen and then lipids were redissolved in 200uL chloroform. 80uL was injected on LC-MS TOF in

positive and negative mode. Chromatograms were then aligned using XCMS in order to identify lipids that were changing. In order to confirm the identity of arachidonic acid, this ion was coeluted with an arachidonic acid standard. Phosphatidic acid identity was confirmed by LC-MS/MS.

RACE pcr:

Race pcr was performed with in the 5' and 3' direction using the GeneRacer kit and the following gene specific primers:

1: tcaaccgcagcgtgtcagaccagttcta

2: ccgcagcgtgtcagaccagttctatcgc

3: tagaactggtctgacacgctgcggttga

4: gcgatagaactggtctgacacgctgcgg

5.7 References

(1) Slavoff, S. A.; Mitchell, A. J.; Schwaid, A. G.; Cabili, M.; Ma, J.; Levin, J. Z.; Budnik, B.; Rinn, J. L.; Saghatelian, A. *Nature Chemical Biology* **2012**, *9*, 59.

(2) Stern-ginossar, N.; Weisburd, B.; Michalski, A.; Le, V. T. K.; Hein, M. Y.; Huang, S.; Ma, M.; Shen, B.; Qian, S.; Hengel, H.; Mann, M.; Ingolia, N. T.; Weissman, J. S. *Science* **2013**, *338*, 1088.

(3) Ingolia, N. T.; Lareau, L. F.; Weissman, J. S. *Cell* **2011**, *147*, 789.

(4) Galindo, I. G.; Pueyo, J. I.; Fouix, S.; Bishop, S. A.; Couso, J. P. *PLoS Biol* **2007**, *5*, 1052.

(5) Kondo, T.; Plaza, S.; Zanet, J.; Benrabah, E.; Valenti, P.; Hashimoto, Y.; Kobayashi, S.; Payre, F.; Kageyama, Y. *Science* **2010**, *329*, 326.

(6) Hashimoto, Y.; Niikura, T.; Tajima, H.; Yasukawa, T.; Sudo, H.; Ito, Y.; Kita, Y.; Kawasumi, M.; Kouyama, K.; Doyu, M.; Sobue, G.; Koide, T.; Tsuji, S.; Lang, J.; Kurokawa, K.; Nishimoto, I. *PNAS* **2001**, *98*, 6336.

(7) Guo, B.; Zhai, D.; Cabezas, E.; Welsh, K.; Nouraini, S.; Satterthwait, A. C.; Reed, J. C. *Nature* **2003**, *423*, 456.

(8) Ikonen, M.; Liu, B.; Hashimoto, Y.; Ma, L.; Lee, K.-W.; Niikura, T.; Nishimoto, I.; Cohen, P. *PNAS* **2003**, *100*, 13042.

(9) Carvunis, A. R.; Rolland, T.; Wapinski, I.; Calderwood, M. A.; Yildirim, M. A.; Simonis, N.; Charloteaux, B.; Hidalgo, C. A.; Barbette, J.; Santhanam, B.; Brar, G. A.; Weissman, J. S.; Regev, A.; Thierry-Mieg, N.; Cusick, M. E.; Vidal, M. *Nature* **2012**, *487*, 370.

(10) Guttman, M.; Amit, I.; Garber, M.; French, C.; Lin, M. F.; Feldser, D.; Huarte, M.; Zuk, O.; Carey, B. W.; Cassady, J. P.; Cabili, M.; Jaenisch, R.; Mikkelsen, T. S.; Jacks, T.; Hacohen, N.; Bernstein, B. E.; Kellis, M.; Regev, A.; Rinn; Lander, E. S. *Nature* **2009**, *458*, 223.

(11) Ashburner, M.; Ball, C. A.; Blake, J. A.; Botstein, D.; Butler, H.; Cherry, J. M.; Davis, A. P.; Dolinski, K.; Dwight, S. S.; Eppig, J. T.; Harris, M. A.; Hill, D. P.; Issel-Tarver, L.; Kasarskis, A.; Lewis, S.; Matese, J. C.; Richardson, J. E.; Ringwald, M.; Rubin, G. M.; Sherlock, G. *Nature Genetics* **2000**, *25*, 25.

(12) Peck, D.; Crawford, E. D.; Ross, K. N.; Stegmaier, K.; Golub, T. R.; Lamb, J. *Genome Biology* **2006**, *7*, R61.

(13) Arcaro, A.; Wymann, M. P. *Biochem. J.* **1993**, *296*, 297.

(14) Stebbins, C. E.; Russo, A. A.; Schneider, C.; Rosen, N.; Hartl, F. U.; Pavletich, N. *Cell* **1997**, *89*, 239.

(15) Marks, P. A.; Breslow, R. *Nature biotechnology* **2007**, *25*, 84.

(16) Yoshida, M.; Kijima, M.; Akita, M.; Beppu, T. *J Biol Chem* **1990**, *265*, 17174.

(17) Subramanian, A.; Tamayo, P.; Mootha, V. K.; Mukherjee, S.; Ebert, B. L.; Gillette, M. A.; Paulovich, A.; Pomeroy, S. L.; Golub, T. R.; Lander, E. S.; Mesirov, J. P. *PNAS* **2005**, *102*, 15545.

(18) Tian, B.; Nowak, D. E.; Jamaluddin, M.; Wang, S.; Brasier, A. R. *Journal of Biological Chemistry* **2005**, *280*, 17435.

(19) Phong, M. S.; Van Horn, R. D.; Li, S.; Tucker-Kellogg, G.; Surana, U.; Ye, X. S. *Mol Cell Biol* **2010**, *30*, 3816.

(20) Kabe, Y.; Ando, K.; Hirao, S.; Yoshida, M.; Handa, H. *Antioxidants and Redox Signaling* **2005**, *7*, 395.

(21) Bubici, C.; Papa, S.; Dean, K.; Franzoso, G. *Oncogene* **2005**, *25*, 6731.

(22) Nathan, C. *Nature* **2002**, *420*, 846.

(23) Dinarello, C. A.; Cannon, J. G.; Wolff, S. M.; Bernheim, H. A.; Beutler, B.; Cerami, A.; Figari, I. S.; Palladino, M. A.; O'Connor, J. V. *Journal of Experimental Medicine* **1986**, *163*, 1433.

(24) Taub, D. D.; Anver, M.; Oppenheim, J. J.; Longo, D. L.; Murphy, W. J. *The Journal of Clinical Investigation* **1996**, *97*, 1931.

(25) Koch, A. E.; Polverini, P. J.; Kunkel, S. L.; Harlow, L. A.; Dipietro, L. A.; Elner, V. M.; Elner, S. G.; Strieter, R. M. *Science* **1992**, *258*, 1798.

(26) Harada, A.; Sekido, N.; Akahoshi, T.; Wada, T.; Mukaida, N.; Matsushimi, K. *Journal of Leukocyte Biology* **1994**, *56*, 559.

(27) Piomelli, D. *Current Opinion in Cell Biology* **1993**, *5*, 274.

(28) Haliday, E. M.; Ramesha, C. S.; Ringold, G. *EMBO* **1991**, *10*, 109.

(29) Irvine, R. F. *Biochem. J.* **1982**, *204*, 3.

(30) Calvo, S. E.; Pagliarini, D. J.; Mootha, V. K. *Proc Natl Acad Sci U S A* **2009**, *106*, 7507.

(31) Parola, A. L.; Kobilka, B. K. *J Biol Chem* **1994**, *269*, 4497.

(32) Jennings, M. D.; Pavitt, G. D. *Nature* **2010**, *465*, 378.

(33) Yoo, H.; Yoo, J. K.; Lee, J.; Lee, D. R.; Ko, J. J.; Oh, S. H.; Choo, Y. K.; Kim, J. K. *Biochemical and Biophysical Research Communications* **2011**, *415*, 567.

(34) Cano, V. S. P.; Jeon, G. A.; Johansson, H. E.; Henderson, C. A.; Park, J.-H.; Valentini, S. R.; Hershey, J. W. B.; Park, M. H. *FEBS J* **2008**, *275*, 44.

(35) Horard, B.; Vanacker, J. M. *J Mol Endocrinol* **2003**, *31*, 349.

(36) Kallen, J.; Schlaeppi, J. M.; Bitsch, F.; Filipuzzi, I.; Schilb, A.; Riou, V.; Graham, A.; Strauss, A.; Geiser, M.; Fournier, B. *J Biol Chem* **2004**, *279*, 49330.

Appendix:

A.1 Experiments to identify bioactive lipids in cancer and inflammation

A.1.1 Discovery of osteoclast secreted lipids that promote bone cancer metastasis.

In bone cancers, osteoclasts have been tied to increased cancer metastasis and tumor growth[1]. Interactions in the tumor microenvironment between osteoclasts and cancer cells can lead to both osteolysis and tumor proliferation[1]. In particular, osteolysis leads to the secretion of bone derived growth factors, such as insulin-like growth factor 1 (IGF1), and transforming growth factor-β (TGF-β), and raises extracellular calcium ion levels. IGF1 and TGF-β are recognized by receptors on the tumor cell and trigger downstream phosphorylation through the MAPK and SMAD pathways triggering growth. Extracellular calcium levels activate the calcium pump and also contribute to tumor growth. In turn, tumor cells secrete parathyroid hormone related peptide (PTHrP), which promotes osteoclast differentiation—completing the positive feedback loop.

Peroxisome proliferator-activated-receptor ϒ (PPARϒ) is also important in osteoclast differentiation and bone resorption[2]. Agonism of PPARϒ with rosiglitazone, a PPARϒ ligand, promotes osteoclastogenesis and breakdown of bone. PPARϒ activation can also lead to changes in lipid metabolism[3].

In addition to the secretion of tumorigenic protein hormones there is also the possibility that osteoclasts secrete pro-tumorigenic lipids. Therefore, osteoclasts were treated with or without rosiglitazone in order to promote osteoclastogenesis, bone resorption, and accentuate lipid production from macrophages[4]. Media secreted from

treated or untreated osteoclast was collected, lipids were extracted, and were analyzed using a global LC-MS approach. This method indicated a number of up and down regulated metabolites (Figure A.1). Coelution with synthetic standards was used to confirm upregulation of polyunsaturated fatty acids including ω-6 arachidonic acid, and eicosopentanoic acid, while MRM analysis confirmed that phosphatidylcholines (PC) were downregulated. Notably both arachidonic and eicosopentanoic acid are proinflammatory metabolite precursors. The effect of these lipids on cancer cell proliferation is being investigated further.

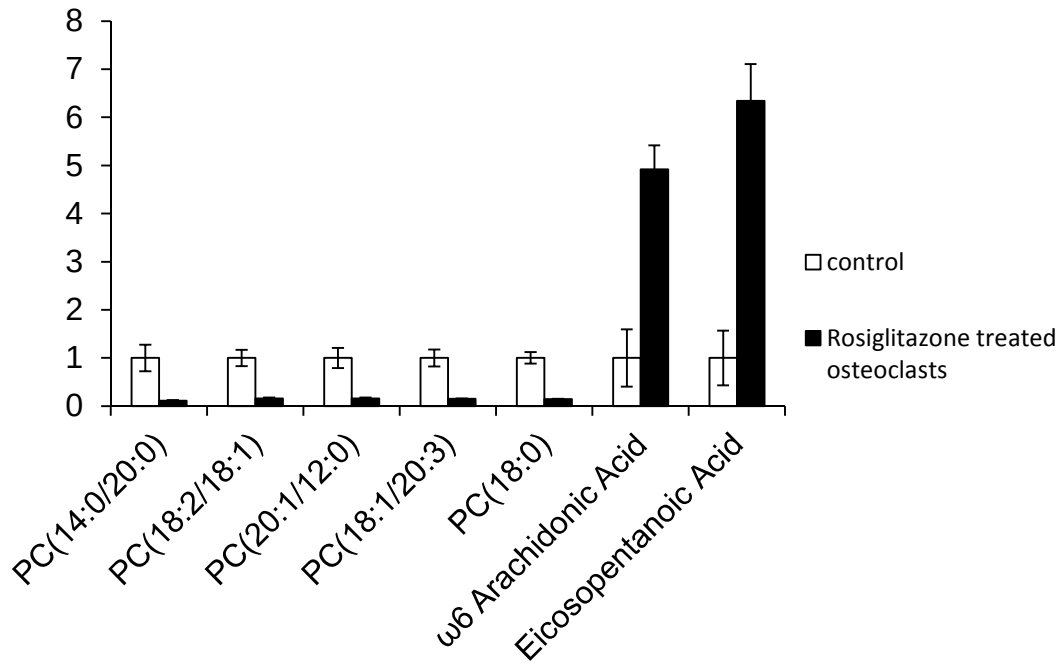


Figure A.1: Lipids up or down regulated in rosiglitazone treated or control osteoclast. ω-6 arachidonic acid and eicosopentanoic acid were verified by coelution. PCs were verified by MRM. Statistics was performed with a Student's t-test, *, $p \leq 0.05$, **, $p \leq 0.01$.

A.1.2 Identification of abnormal lipids levels in alopecia inducing mouse milk

Adiponectin is a cytokine secreted principally by adipocytes that has been shown to regulate lipid metabolism and homeostasis[5]. Recently it was found that knock-out

mice lacking adiponectin, or transgenic mice overexpressing adiponectin produced milk that induced alopecia in pups. Lipids, and in turn adiponectin, can have proinflammatory functions, and in order to determine how effecting adiponectin levels was triggering alopecia, lipids in knock out (KO), transgenic (TG), and wild type (WT) mouse milk were examined[6]. A global lipid profiling approach revealed that adiponectin KO mouse milk lacked lysoPCs, relative to WT milk (Figure A.2 and A.3). Additionally, neutral loss mass spectrometry of triglyceride species revealed that certain triglycerides were depleted in KO and TG mice. The role of these lipids in alopecia merits further study.

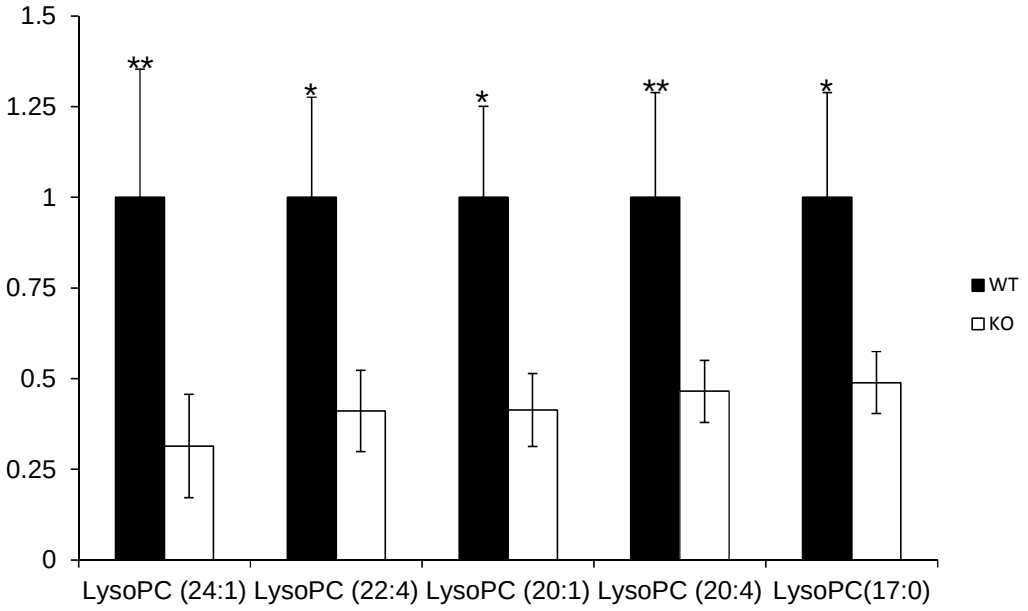


Figure A.2: LysoPCs are downregulated in the milk of adiponectin KO mice. Changes in lysoPCs were identified using a global lipidomics approach. The identity of some lysoPCs was further validated by MRM. Statistics was performed with a Student's t-test, *, $p \leq 0.05$.

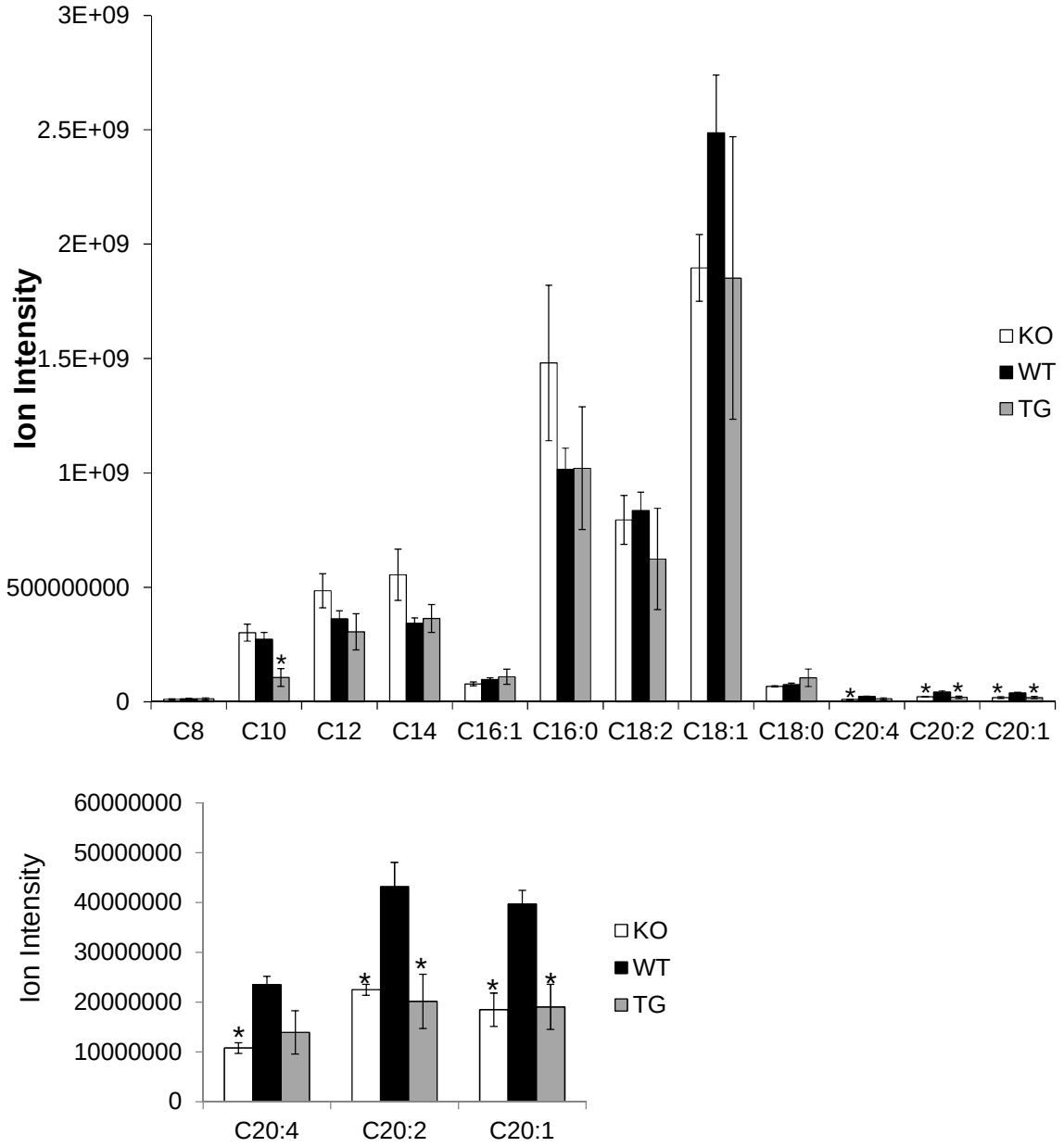


Figure A.3: Triglyceride levles in adiponectin KO and TG mouse milk. (Top) Although the most abundant triglycerides are unchanged in KO and TG milk, several less abundant triglycerides are depleted compared to WT. (Bottom) Close up of changes in C20:4, C20:2, and C20:1 triglycerides illustrates the depletion of these triglycerides relative to the WT sample. Statistics was performed with a Student's t-test, *, $p \leq 0.05$

Table A.1: Complete list of SEP detected peptides and validation methods used to confirm them from chapter 3.

validation						
imaging	heavy/light tryptic peptide synthesis	PAGE size identification	Full Length coelution			Detected Peptide
X					refseq	AAPGALPEAAVGPR
					refseq	AGAPAVGLLLANER
X					refseq	QLPPAAAVGDAGQLGR
					refseq	
		X			refseq	ATPGLQQHQQPPGPGR
					refseq	
					refseq	
					refseq	
X		X			refseq	IVVDELSNLK
					refseq	
					refseq	
	X				refseq	
					refseq	TAPSSTATTASASCAATR
X	X	X			refseq	LQVGPADTQPR
X					refseq	STAACQTSSIATR
					refseq	GSSAAVGPR
X					refseq	TAAAAAAGTITRPR
		X			refseq	GVGGQAALFAAGR
					refseq	
					refseq	
					refseq	
		X			refseq	AVAAAAAAAPDPGGR
					refseq	
X		X			refseq	GGLGAASIAADGAPR
X					refseq	SSTPAPPQGQFLPPSI
X		X			refseq	VAVEEGLPGDPVAER
					refseq	
					refseq	EGSVHPQVE
					refseq	GAIGGGGAGVQGQTAGAR
					refseq	VAAVAVGSQAVLQILSR
					refseq	WTSSTSSPNTSGAPR
					refseq	NPPLVQDTVSGK
					refseq	QTAFGKWYESLLNNR
					refseq	AVAGAAAGAGGR
X					refseq	AEEQPGLGPGAAGR
					refseq	RAVPAQGLLQSTPTCMPWTP
	X				refseq	NTTQESLEKGP
					refseq	EALNEFLTR
	X				refseq	AEPLQTAGQAGR
					refseq	AGNLILLQ
				X	refseq	STTIGGMNQR
					refseq	ERPANSLIDQCSQR
					refseq	VFFKNLLAFAR
	X				refseq	AELSFLNR
		X			refseq	LLPLGASPAGVVGGGLAPPR

(Continued) Table A.1: Complete list of SEP detected peptides and validation methods used to confirm them from chapter 3.

			refseq	
			refseq	
			refseq	TWLPSCEDLTLPGGR
X			refseq	FLPVDLSLLR
			refseq	SLSSYGACSR
	X		refseq	GPSGTQEMGPLSR
			refseq	
				APEPGAVLAPAEVVLR
				LLVSGSPSAETLPLR
				ALAQGSLTPSQIYSA
				LSAPQPGPDILQAPAR
				VYIFQPVFEQYAK
				NEQTELLYNK
				ILEDFLPPSSSRPQS
				DLPGVAPPRPSLSLSGP
				AAASGQPRPEMQCPAEQTEIK
				AQHGVHSNTASPGLPAGAPR
	X			KQGGFVQVSANAL
				LNINQSIAVSTATQR
				LPGQATTQQTFDQR
				LVSAVLAGKE
				PAVAAATLHLPAAPEGPH
				QELIGASLHTAR
			lincRNA	THLGTEGQCDLPGAGGPAR
				TSDAPRPSATPPGADPLNSAGPGAR
				VTSWDGQNPPR
				AAPGPTAAAAAQASAAAR
				RLLIPPEK
				SPTTDSYGIPQGCK
				DYILSLEMFSILLWG
	X			HGHSFPDPGLLLQNQGD
				HDASSSPLGPPR
				CLVYVLDLITDACTIKPLFNK
				ASPGEAGPAGGAAAGQGAPR
				GAWGGGQLATAGSGPGQR
			lincRNA	DTEVLINTMSK
				VYKWLLCNVE
				CPFVLLMSSMILLR
				KPVFLLLLSIR
				FIPTEAWYSAGR
		X		QVLITNKNQ
			lincRNA	IKFLLAPEENK
			lincRNA	QRIPCVVILTK
			lincRNA	KTLPMMGMIR
			lincRNA	QVNEETLK
			lincRNA	KNLFQNTSR
			lincRNA	RAGYSELE

(Continued) Table A.1: Complete list of SEP detected peptides and validation methods used to confirm them from chapter 3.

	QMSSNILK
	VAHENYMKFK
	GIALGDIPNAR
	VLLDQHQR
	YYELQRGTR
	GEMERGEIK
	CQDILEAGKR
	DLGSPMLK
	TASPYSRPE
	LTVAGQGR
	SPFWAGQGQSR
	NLAGGSGLIP
	AAALQFDLR

Table A.2: Complete list of detected peptides from SEPs identified in chapter 3 along with other detected peptides that map to the same SEP. Detected peptides listed in red do not meet the scoring criteria to identify a SEP on their own, but when matched with a group of other detected peptides occuring from the same sORF contribute to the confidence of the SEP identification.

Detected Peptide	Additional Detected Peptides from same sORF	PAGE ID
AAPGALPEAAVGPR		
AGAPAVGLLLANER		
QLPPAAAVGDAGQLGR		
	APGGAAAGPGAPGCGGAGGQGPAPGGAAAAAAR	
ATPGLQQHQQPPGPGR		
	ATPPGGTGHEGLSGGAADVASGVGSGR	ATPPGGTGHEGLSGGAADVASGVGSGR
		GMTDSPPPGHPEEK
		HRWPPPPGGAAPAPVR
IVVDELSNLK		IVVDELSNLKK
	QQQNSNIFFLADR	QQQNSNIFFLADR
	NILDELKK	
	EYQEIENLDK	EYQEIENLDKTK
TAPSSTATTASASCAATR		
LQVGPADTQPR		LQVGPADTQPR
STAACQTSSIATR		
GSSAAVGPR		
TAAAAAAGTITRPR		
GVGGQAALFAAGR		GVGGQAALFAAGR
	AGGDLPLQPQPGGAAAR	AGGDLPLQPQPGGAAAR
	AAQAFFPAAELAQAGPER	AAQAFFPAAELAQAGPER
		AAHPHHAQVHPAVALQPAR
AVAAAAAAAPDPGGR		AVAAAAAAAPDPGGR
		GCESAAAEAAAAEEAAAGGGVGEPAPGRR
GGLGAASIAADGAPR		GGLGAASIAADGAPR
SSTPAPPQGQFLPPSI		
VAVEEGLPGDPVAER		
		DAEQEEEVQR
EGSVHPQVE		
GAIGGGGAGVQGQTAGAR		
VAAVAVGSQAVLQILSR		
WTSSTSSPNTSGAPR		
NPPLVQDTVSGK		
QTAFGKWYESLLNNR		
AVAGAAAGAGGR		
AEEQPGLGPGAAGR		
RAVPAQGLLQSTPTCMPWTP		
NTTQESLEKGP		
EALNEFLTR		
AEPLQTAGQAGR		
AGNLILLQ		
STTIGGMNQR		
ERPANSLIDQCSQR		
VFFKNLLAFAR		
AELSFLNR		
LLPLGASPAGVVGGGLAPPR		LLPLGASPAGVVGGGLAPPR
		QGPKADSDSDLETEGAR
		ADSDSDLETEGAR

(Continued) Table A.2: Complete list of detected peptides from SEPs identified in chapter 3 along with other detected peptides that map to the same SEP. Detected peptides listed in red do not meet the scoring criteria to identify a SEP on their own, but when matched with a group of other detected peptides occuring from the same sORF contribute to the confidence of the SEP identification.

TWLPSCEDLTLPGGR	
FLPVDLSLLR	
SLSSYGACSR	
GPSGTQEMGPLSR	
	GADGGGGAGSAGQIQR
APEPGAVLAPAEVVLR	
LLVSGSPSAETLPLR	
ALAQGSLTPSQIYSA	
LSAPQPGPDILQAPAR	
VYIFQPVFEQYAK	
NEQTELLYNK	
ILEDFLPPSSSRPQS	
DLPGVAPPRPSLSLSGP	
AAASGQPRPEMQCPAEQTEIK	
AQHGVHSNTASPGLPAGAPR	
KQGGFVQVSANAL	
	SETALLALDRPLLPPALR
LNINQSIAVSTATQR	
LPGQATTQQTFDQR	
LVSAVLAGKE	
PAVAAATLHLPAAPEGPH	
QELIGASLHTAR	
THLGTEGQCDLPGAGGPAR	
TSDAPRPSATPPGADPLNSAGPGAR	
VTSWDGQNPPR	
AAPGPTAAAAAQASAAAR	
RLLIPPEK	
SPTTDSYGIPQGCK	
DYILSLEMFSILLWG	
HGHSFPDPGLLLQNQGD	HGHSFPDPGLLLQNQGD
	GGADQNNVQHQPPGEPEGVHQQSASPGGLHDQR
HDASSSPLGPPR	
CLVYVLDLITDACTIKPLFNK	
ASPGEAGPAGGAAAGQGAPR	
GAWGGGQLATAGSGPGQR	
DTEVLINTMSK	
VYKWLLCNVE	
CPFVLLMSSMILLR	
KPVFLLLLSIR	
FIPTEAWYSAGR	
QVLITNKNQ	
IKFLLAPEENK	
QRIPCVVILTK	
KTLPMMGMIR	
QVNEETLK	
KNLFQNTSR	
RAGYSELE	
QMSSNILK	

(Continued) Table A.2: Complete list of detected peptides from SEPs identified in chapter 3 along with other detected peptides that map to the same SEP. Detected peptides listed in red do not meet the scoring criteria to identify a SEP on their own, but when matched with a group of other detected peptides occuring from the same sORF contribute to the confidence of the SEP identification.

VAHENYMKFK
GIALGDIPNAR
VLLDQHQR
YYELQRGTR
GEMERGEIK
CQDILEAGKR
DLGSPMLK
TASPYSRPE
LTVAGQGR
SPFWAGQGQSR
NLAGGSGLIP
AAALQFDLR

Table A.3: List of detected peptides from SEPs identified in chapter 3 along with start codons, SEP length and Chromosome coordinates.

Detected Peptide	Start Codon	SEP length (aa)	Chromosome coordinates
AAPGALPEAAVGPR	ATG	96	chr9:139256352-139264369 strand=-
AGAPAVGLLLANER	GTG	39	chrX:16859470-16888534 strand=-
QLPPAAAVGDAGQLGR	ACG	103	chr10:99092201-99094454 strand=-
ATPGLQQHQQPPGPGR	ATG	83	chr9:139557366-139565706 strand=+
IVVDELSNLK	ATG	96	chr2:190526195-190535440 strand=+
TAPSSTATTASASCAATR	ATG	62	chr7:150646657-150675423 strand=-
LQVGPADTQPR	ATG	88	chr9:123612077-123639492 strand=-
STAACQTSSIATR	ATG	97	chr14:103800538-103809402 strand=-
GSSAAVGPR	stop	78	chr16:89574827-89607413 strand=+
TAAAAAAGTITRPR	GTG	102	chr8:64080459-64125260 strand=+
GVGGQAALFAAGR	GTG	88	chr8:144897399-144897840 strand=-
AVAAAAAAAPDPGGR	acg	91	chr10:98288128-98346562 strand=-
GGLGAASIAADGAPR	ctg	115	chr4:122737616-122745077 strand=-
SSTPAPPQGQFLPPSI	acg	74	chr7:100464771-100471014 strand=+
VAVEEGLPGDPVAER	acg	107	chr11:65686750-65689023 strand=+
EGSVHPQVE	atg	87	chr10:101992055-102005758 strand=-
GAIGGGGAGVQGQTAGAR	atg	143	chr5:180650039-180662529 strand=+
VAAVAVGSQAVLQILSR	atg	77	chr19:42713286-42721897 strand=-
WTSSTSSPNTSGAPR	atg	77	chr19:12949331-12969791 strand=+
NPPLVQDTVSGK	atg	111	chr1:150522391-150532570 strand=+
QTAFGKWYESLLNNR	stop	63	chr3:193363602-193386115 strand=+
AVAGAAAGAGGR	atg	73	chr19:13059508-13067950 strand=-
AEEQPGLGPGAAGR	atg	149	chr7:100032962-100034242 strand=-
RAVPAQGLLQSTPTCMPWTP	atg	54	chr1:160061156-160064154 strand=-
NTTQESLEKGP	stop	32	chr22:41740383-41756157 strand=+
EALNEFLTR	stop	22	chr4:169908762-169911558 strand=-
AEPLQTAGQAGR	atg	59	chr11:118964597-118966163 strand=-
AGNLILLQ	stop	23	chr3:124945640-125042272 strand=-
STTIGGMNQR	atg	26	chr12:48732236-48745011 strand=-
ERPANSLIDQCSQR	atg	54	chr2:131130309-131132956 strand=+
VFFKNLLAFAR	stop	22	chr6:80194734-80199064 strand=-

(Continued) Table A.3: List of detected peptides from SEPs identified in chapter 3 along with start codons, SEP length and chromosome coordinates.

TWLPSCEDLTLPGGR	atg	50	chr7:158799724-158814542 strand=+
FLPVDLSLLR	atg	90	chr1:155532795-155708399 strand=+
SLSSYGACSR	stop	71	chr17:35441928-35444379 strand=+
GPSGTQEMGPLSR	**atg**	102	chr19:3610043-3626771 strand=-
APEPGAVLAPAEVVLR	agg	56	chr22:47048295-47073068 strand=+
LLVSGSPSAETLPLR	atg	128	chr5:34914296-34925392 strand=+
ALAQGSLTPSQIYSA	aag	52	chr22:17092426-17095991 strand=+
LSAPQPGPDILQAPAR	GTG	89	chr19:54693858-54697432 strand=+
VYIFQPVFEQYAK	atg	54	chr15:55609385-55613829 strand=-
NEQTELLYNK	stop	18	chr12:118649944-118650075 strand=
ILEDFLPPSSSRPQS	stop	42	chr2:85132483-85133801 strand=+
DLPGVAPPRPSLSLSGP	atg	65	chr9:130209955-130216851 strand=-
AAASGQPRPEMQCPAEQTEIK	atg	58	chr5:14664778-14699800 strand=+
AQHGVHSNTASPGLPAGAPR	agg	66	chr7:150778180-150780257 strand=-
KQGGFVQVSANAL	atg	136	chr22:32014633-32026837 strand=-
LNINQSIAVSTATQR	AGG	55	chr2:200322928-200323580 strand=+
LPGQATTQQTFDQR	stop	54	chr19:56165091-56185542 strand=+
LVSAVLAGKE	CTG	43	chr1:7863564-7864928 strand=-
PAVAAATLHLPAAPEGPH	atg	49	chr7:100169852-100183655 strand=-
QELIGASLHTAR	stop	119	chr1:228544743-228549628 strand=+
THLGTEGQCDLPGAGGPAR	stop	100	chr10:11925853-11937442 strand=
TSDAPRPSATPPGADPLNSAGPGAR	stop	103	chr19:55737961-55770381 strand=-
VTSWDGQNPPR	ATG	50	chr12:12966292-12982891 strand=+
AAPGPTAAAAAQASAAAR	CTG	108	chr2:231577583-231685792 strand=+
RLLIPPEK	stop	45	chrX:5214450-5216144 strand=+
SPTTDSYGIPQGCK	stop	40	chr1:175913973-176153786 strand=-
DYILSLEMFSILLWG	stop	33	chr3:88101102-88108113 strand=-
HGHSFPDPGLLLQNQGD	stop	122	chr7:66386236-66423532 strand=+
HDASSSPLGPPR	stop	55	chr16:87435666-87438903 strand=-
CLVYVLDLITDACTIKPLFNK	stop	43	chr9:130128866-130129660 strand=+
ASPGEAGPAGGAAAGQGAPR	stop	73	chr1:16905808-16970994 strand=-
GAWGGGQLATAGSGPGQR	ATG	70	chr17:62205639-62207524 strand=-
DTEVLINTMSK	ATT	27	chr1:4036227-4073316 strand=+
VYKWLLCNVE	ATG	41	chr1:157243513-157253900 strand=+
CPFVLLMSSMILLR	STOP	33	chr10:119806332-119859641 strand=+
KPVFLLLLSIR	STOP	32	chr11:3532972-3542051 strand=+
FIPTEAWYSAGR	ATG	86	chr11:82783129-82805398 strand=+
QVLITNKNQ	ATG	29	chr11:65266565-65274602 strand=-
IKFLLAPEENK	ATG	43	chr16:3054772-3058645 strand=+
QRIPCVVILTK	stop	73	chr19:23278060-23286908 strand=+
KTLPMMGMIR	stop	30	chr2:107137814-107160732 strand=+
QVNEETLK	stop	143	chr3:107852804-107857456 strand=+
KNLFQNTSR	stop	59	chr4:10069715-10074643 strand=-
RAGYSELE	ATG	69	chr7:96251318-96293650 strand=-
QMSSNILK	stop	50	chr15:31008518-31061502 strand=+
VAHENYMKFK	stop	59	chr21:35345400-35353552 strand=+
GIALGDIPNAR	GTG	18	chr6:68590370-68642035 strand=+
VLLDQHQR	stop	23	chr6:141167131-141219546 strand=-

(Continued) Table A.3: List of detected peptides from SEPs identified in chapter 3 along with start codons, SEP length and chromosome coordinates.

YYELQRGTR	AAG	43	chr15:59060273-59063173 strand=-
GEMERGEIK	ATG	18	chr17:41373439-41383338 strand=-
CQDILEAGKR	ATC	70	chr19:23441500-23457032 strand=-
DLGSPMLK	ATG	52	chr2:23598100-23604170 strand=-
TASPYSRPE	ATG	58	chr2:66653867-66660602 strand=-
LTVAGQGR	ATG	66	chr20:4173737-4176599 strand=+
SPFWAGQGQSR	stop	101	chrX:118425492-118469573 strand=+
NLAGGSGLIP	stop	41	chrX:1515320-1517852 strand=-
AAALQFDLR	stop	23	chr21-35303432-35308177 strand=+

A.2 Works Cited

(1) Mundry, G. R. *Nature Reviews Cancer* **2002**, *Metastasis to Bone: Causes, Consequences and Therapeutic Opportunities*, 584.

(2) Wan, Y.; Chong, L.-W.; Evans, R. M. *Nature Medicine* **2007**, *13*, 1496.

(3) Chinetti, G.; Fruchart, J. C.; Staels, B. *Inflammation Research* **2000**, *49*, 497.

(4) Chawla, A.; Barak, Y.; Nagy, L.; Liao, D.; Tontonoz, P.; Evans, R. M. *Nature Medicine* **2001**, *7*, 48.

(5) Berg, A. H.; Combs, T. P.; Scherer, P. E. *Trends Endocrinol Metab* **2002**, *13*, 84.

(6) Ouchi, N.; Parker, J. L.; Lugus, J. J.; Walsh, K. *Nature Reviews Immunology* **2011**, *11*, 85.